FE Project Midterm Report Draft

Sponsor: Richard Yeh, Cantor Fitzgerald.

Team members: Zhenyu Chen, Jiongjia Fang, Yilun Qian, Jia Xu, Kuangdi Zheng.



目录

Part one: predicting price changes with signed order imbalance for equity securities 2

Appendix I 13

Part Two: where stocks trade? 15

Attachment II 41

## Part one

***Predicting price changes with Signed Order Imbalance for equity securities***

**Objective:**

This project aims to improve one period ahead price prediction using order imbalance information from the current time period.

**Background information:**

**Motivation behind the project:**

Securities price prediction is a golden grail. The ability to accurately predict price changes lead to direct profits and revenue generation for the firm. Endless efforts have been devoted into coming up with great models, and many are fruitless while others are more promising. The most successful strategies are usually very secretive and involve advanced mathematical modeling. In this paper, the team dabbled at building one simple model based on signed order imbalances, an idea inspired by Easley, Lopez de Prado and O’Hara’s paper on “Flow Toxicity and Liquidity in a High Frequency World”.

**Definition of SOI:**

**Explanation of the definition:**

defines the total volume of a bucket. Such bucket volume needs to identical throughout the day. volume(i, i-1) defines the total volume in a singular “bin” within each bucket. This term need not to be fixed, although it is possible to fix such a term based on volume (i.e. fixed bin implementation). Alternatively, one can define this term in terms of time; to put it in another way, we aggregate all the volumes within a fixed time interval and define it as volume(i,i-1), where i is the time index associated with each time interval. b(i) is a discrete function only takes value {+1,-1,0}, which indicates the act (Bull/Sell/Neutral) of the corresponding bin.

The classification method is explained in detail in the following part.

**Buy/Sell/Neutral Classification (i.e. b(·) Calculation ):**

Multiple buy/sell/neutral classification schemes are explored:

1. Lee-Ready type classification based solely on trade data. In this algorithm, transactional level prices for trades are used. A trade is classified as a buy trade if the transactional price at time t is higher than the price at t-1, and classified as a sell trade if the transactional price at time t is lower than the price at t-1. In case both transactional prices are identical, we look back one period (i.e. t-2 to t-1 period) and follow the same procedure. The rationale behind this is in case of the tie, we implicitly assume the momentum will continue. If the t-2 to t-1 period return is still 0, then we classify the bin of t-1 to t as neutral, and return 0. (i.e. this particular trade has no weight in calculating signed order imbalance)
2. Nearest quotes with artificial delay based classification.

Let us first explain how we use the nearest quotes to classify trades, then we will explain the rational behind adding a constant artificial delay time to all the quotes.

For each trade, this algorithm looks at the nearest available quote ahead of it (in terms of exchange time). If the transacted price is closer to the bid of the quote, we classify the trade as a sell by setting b to -1 (i.e. we are taking the bid). Alternatively if the transacted price is closer to the ask of the quote, the trade is classified as a buy (or b = 1). If the transacted price is equidistant, we assigned the trade as neutral.

The trades reported to the consolidates tape is usually late than the ), to To remedy for that case, we introduce random artificial delay to the quotes and follow the same algorithm for classification.

Quotes delay technique can and has been applied to the raw data for this project even if we do not use technique 2) for classification.

1. The third classification scheme involves calculating time weighted EMAs of all the quotes, and starting with the nearest EMA quote, we follow technique 2) classification.

**Analysis:**

Before proceeding to building an actual predictive model, it is important to first ascertain our intuition that SOI is well correlated with price returns. The simple linear regression model we used to verify is as follows:

,

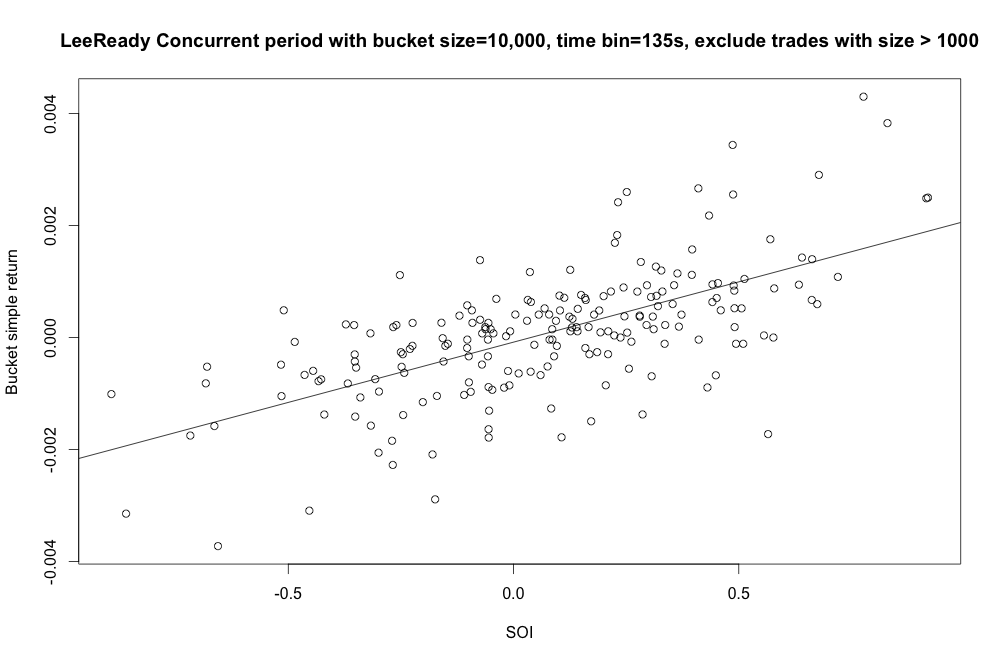
There are several free model parameters we need to determine in order to calculate SOI. The two major ones are bucket size and bin size. The first one is determined by volume, while the second one is determined by time in our implementation, though both can be determined either by volume or time. It is important to calibrate the ideal bucket and bin size, because the regression results are sensitive to the bucket size and bin size.

To illustrate why this can be the case, imagine a bucket size of 10 shares, a 100 shares trade would fill all 10 consecutive buckets, and the computed SOI for all the ten periods would be 1. From the surface, it is almost as if we can detect a pattern of autocorrelation of SOI measures, yet the consecutive SOI measures are simply due to splitting the trades. As for time bins, a single time bin aggregates all the trades within the time period, and is consider the smallest unit of processing. If the time bin is set too small, we may be including a lot of ultra-short term noise, if the time bin is too large, we lose a level of granularity, which can be helpful in improving the model.

To help establishing feasibility, we used AMZN stock for 4/23/2013. By using brute computing force, and regressing all the possible bucket size, bin size, and block trade threshold limit within a pre-specified limits, we find the following combinations to be ideal in terms of adjusted R^2.

|  |  |
| --- | --- |
| **Bucket(shares)\_bin(s)\_Exclude(shares)** | **Adj-R^2** |
| 10000 \_ 135 \_ 1000 | 41.30% |
| 10000 \_ 150 \_ 1000 | 41.30% |
| 10000 \_ 120 \_ 1000 | 37.40% |
| 8000 \_ 135 \_ 1000 | 37.10% |
| 9000 \_ 120 \_ 1000 | 36.60% |

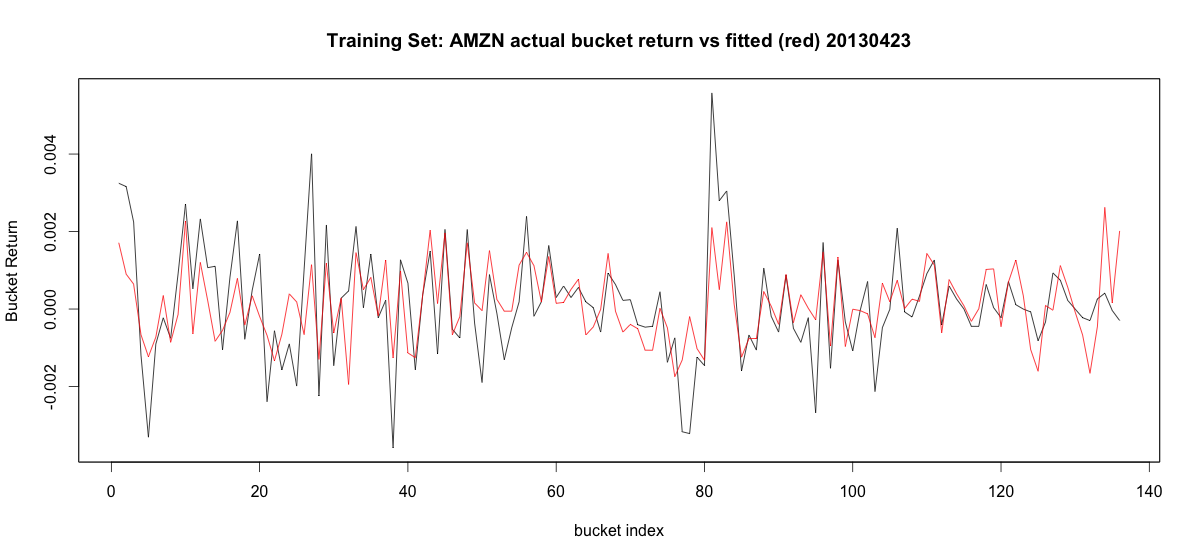
To give the readers an idea how well a R^2 of 41.30% look like, the regression plot is shown below:

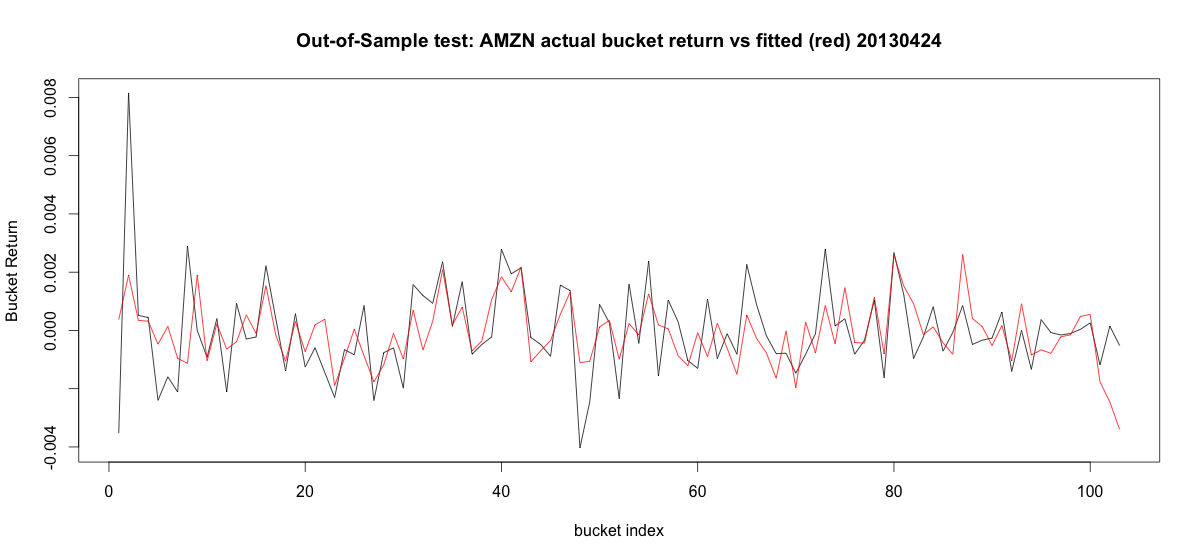


After achieving this step, we moved on to 1) regressing the SOI and the cross term against trade returns and 2) adjusting for the latencies from exchanges, the results improved to:

|  |  |
| --- | --- |
| **Bucket(shares)\_bin(s)\_Exclude(shares)** | **Adj-R^2** |
| 2000 \_ 90 \_ 2000 | 52.18% |
| 15000 \_ 135 \_ 1000 | 50.03% |
| 2000 \_ 105 \_ 2000 | 49.68% |
| 14000 \_ 135 \_ 1000 | 47.79% |
| 15000 \_ 105 \_ 1000 | 47.29% |

We then tested the model robustness across time. Use AMZN 2013/04/23rd data as training data, and 24th data as test set.





Then model captures the movement direction quite well, it correctly predict 94.2% of the price movement direction on 24th. This ratio is calculated as . But as the plots indicated, it is not strong in terms of magnitude depiction.

Next, we want to assess the marginal effect by using different types of classification rules, and the one we are specifically interested is buy/sell classification rule #2, and the weaker R^2 results demonstrated below has shown the Lee-Ready type of classification rule is superior:

|  |  |
| --- | --- |
| **Bucket(shares)\_bin(s)\_Exclude(shares)** | **Adj-R^2** |
| 10000 \_ 60 \_ 3000 | 12.5% |
| 6000 \_ 30 \_ 1000 | 11.4% |
| 6000 \_ 150 \_ 1000 | 9.5% |
| 10000 \_ 60 \_ 5000 | 9.5% |
| 5000 \_ 60 \_ 1000 | 9.5% |

Thus far, we have established the usefulness of using a cross term, the superiority of using delayed exchange time over arrival time, and the validity of SOI’s correlation with price movement.

Thus far, we have used trades’ transactional prices to compute bucket returns. Yet sometimes, practitioners are more interested to predict the quote changes. Not only trade level price changes are noise-prone, one is not guaranteed to be executed at the recorded trade price. So instead, we reformulated the problem, and define mid quote changes in between bucket as price returns.

Next, we assess the We first calculated EMA of all the quotes, and regress the exponential moving average of mid quotes against SOI multiplied by the variance of trade prices within a bucket:

The EMA calculation is defined as:

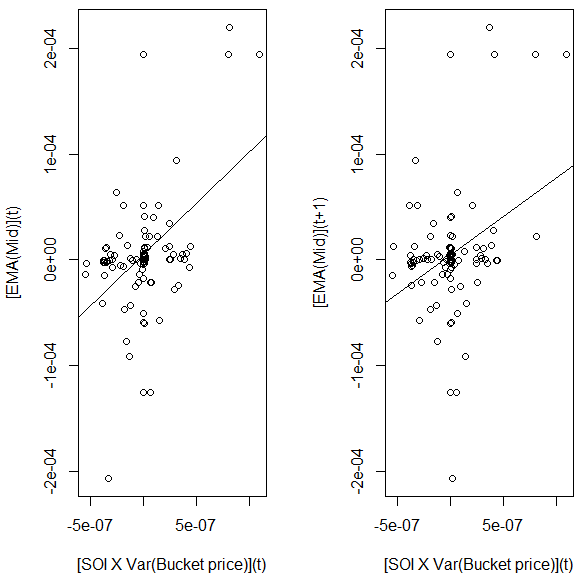
In addition, we have predicted one period ahead R(t) using the same model:

This model is useful in predicting the one period ahead mid-quote. We also introduce one additional twist by artificially delaying the exchange time on the quotes.

We tested this model using five different stocks for five consecutive days. As the stock Below are the output we obtained for one stock AGN, they left column head contains the parameters specific to each regression run, namely the quotes delay, and the EMA decay

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **4/23/2013** |  | **4/24/2013** |  | **4/25/2013** |  |
| *Concurrent Regression* | | | | | |
| Key (delay(s)\_decay) | Adj R2 | Keys | Adj R2 | Keys | Adj R2 |
| 0.025\_1.5 | 33.83% | 0\_0.5 | 20.58% | 0.055\_0.5 | 24.58% |
| 0.03\_1.5 | 33.80% | 0.1\_0.5 | 18.39% | 0.075\_0.5 | 24.28% |
| 0.025\_1.25 | 33.70% | 0.075\_0.5 | 18.18% | 0.08\_0.5 | 24.18% |
| 0.03\_1.25 | 33.70% | 0.09\_0.5 | 18.08% | 0.09\_0.5 | 24.12% |
| 0.03\_1 | 33.19% | 0.085\_0.5 | 18.08% | 0.085\_0.5 | 24.12% |
| *One Step Ahead prediction* | | | | | |
| Key (delay(s)\_decay) | Adj R2 | Keys | Adj R2 | Keys | Adj R2 |
| 0.025\_1.5 | 42.25% | 0\_0.5 | 14.54% | 0.055\_0.5 | 4.64% |
| 0.03\_1.5 | 42.12% | 0\_0.75 | 13.95% | 0.005\_0.5 | 4.56% |
| 0.025\_1.25 | 40.88% | 0\_1 | 13.02% | 0.09\_0.5 | 4.49% |
| 0.03\_1.25 | 40.76% | 0\_1.25 | 11.34% | 0.075\_0.5 | 4.44% |
| 0.025\_1 | 38.90% | 0\_1.5 | 9.25% | 0.05\_0.5 | 4.39% |

|  |  |  |  |
| --- | --- | --- | --- |
| **4/26/2013** |  | **4/29/2013** |  |
| *Concurrent Regression* | | | |
| Key | Adj R2 | Key | Adj R2 |
| 0.06\_0.5 | 14.66% | 0.045\_0.5 | 43.07% |
| 0.06\_0.75 | 13.98% | 0.085\_0.5 | 42.20% |
| 0\_0.5 | 11.72% | 0.07\_0.5 | 42.00% |
| 0.06\_1 | 11.54% | 0.065\_0.5 | 41.77% |
| 0.085\_0.5 | 9.70% | 0.06\_0.5 | 41.74% |
| *One Step Ahead prediction* | | | |
| Key | Adj R2 | Key | Adj R2 |
| 0.1\_0.5 | 7.25% | 0.005\_0.5 | 1.30% |
| 0.07\_0.5 | 7.19% | 0\_0.5 | 0.80% |
| 0.095\_0.5 | 7.18% | 0.075\_0.5 | 0.51% |
| 0.09\_0.5 | 7.14% | 0\_1.5 | 0.45% |
| 0.075\_0.5 | 7.08% | 0\_1.25 | 0.24% |



Above are two sample plots correspond to the cells high-lighted in cyan.

We also did one out of sample run by first estimating the delay and decay parameters by averaging these parameter values from 4/23 to 4/26. Then we obtained the 4/29 concurrent and one period ahead predictions using the estimated model parameters (set delay = 0.035s, and decay = 0.75). Lastly we compare the R^2s with in sample calibrated parameters for 4/29. Below are the differences in performance:

|  |  |  |  |
| --- | --- | --- | --- |
| Key | Out of Sample Concurrent R2 | Key | Best In Sample Concurrent R2 |
| 0.035\_0.75 | 35.54% | 0.035\_0.75 | 43.07% |
| Key | Out of Sample Prediction R2 |  | Best In Sample Prediction R2 |
| 0.045\_0.5 | 0.00% | 0.005\_0.5 | 1.30% |

Even though the decay and delay parameters are not very stable from day to day, using an average value of these parameters to construct out of sample regression yields satisfactory results for this particular example.

Below are the results we have obtained using the same methodology for AMZN stocks.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **4/23/2013** |  | **4/24/2013** |  | **4/25/2013** |  |
| *Concurrent Regression* | | | | | |
| Key (delay(s)\_decay) | Adj R2 | Key | Adj R2 | Key | Adj R2 |
| 0.045\_0.75 | 29.91% | 0.03\_0.75 | 6.50% | 0.005\_1.5 | 4.20% |
| 0.045\_0.5 | 29.43% | 0.035\_0.75 | 6.49% | 0.005\_1.25 | 4.11% |
| 0.045\_1 | 26.65% | 0.06\_0.5 | 6.46% | 0.005\_1 | 3.69% |
| 0.045\_1.25 | 14.90% | 0.055\_0.5 | 6.45% | 0.005\_0.75 | 2.86% |
| 0.005\_0.75 | 3.70% | 0.05\_0.5 | 6.38% | 0.095\_1.5 | 2.79% |
| *One Step Ahead Prediction* | | | | | |
| Key | Adj R2 | Key | Adj R2 | Key | Adj R2 |
| 0.085\_1.5 | 19.34% | 0.055\_0.5 | 4.49% | 0.005\_1.25 | 8.22% |
| 0.085\_1.25 | 19.32% | 0.05\_0.5 | 4.46% | 0.005\_1.5 | 8.02% |
| 0.085\_1 | 18.19% | 0.06\_0.5 | 4.44% | 0.005\_1 | 7.82% |
| 0.085\_0.75 | 16.00% | 0.065\_0.5 | 2.54% | 0.095\_0.75 | 7.28% |
| 0.085\_0.5 | 12.65% | 0.1\_0.5 | 2.14% | 0.1\_0.75 | 7.20% |

|  |  |  |  |
| --- | --- | --- | --- |
| **4/26/2013** |  | **4/29/2013** |  |
| *Concurrent Regression* | | | |
| Key | Adj R2 | Key | Adj R2 |
| 0\_0.5 | 20.73% | 0.04\_0.75 | 16.67% |
| 0\_0.75 | 19.48% | 0.045\_0.75 | 16.48% |
| 0\_1 | 16.31% | 0.02\_1 | 16.39% |
| 0\_1.25 | 13.14% | 0.04\_1 | 15.89% |
| 0.055\_0.75 | 11.40% | 0.045\_1 | 15.62% |
| *One Step Ahead Prediction* | | | |
| Key | Adj R2 | Key | Adj R2 |
| 0\_0.5 | 9.07% | 0.04\_0.75 | 19.54% |
| 0.05\_0.5 | 8.99% | 0.045\_0.75 | 19.35% |
| 0.05\_0.75 | 8.35% | 0.04\_1 | 18.69% |
| 0.09\_0.5 | 7.51% | 0.045\_1 | 18.42% |
| 0.09\_0.75 | 7.42% | 0.04\_0.5 | 17.71% |

Results on BA, CAT, TIF are in the appendix.

We see the results are particularly good for 4/23/2013, which is the date when White House’s twitter account was hacked. As a result, we expect SOI in generally to work better when the market is relatively more volatile and filled with more impactful news.

We believe in this intuition, because as more market participants are aware of potential impactful news, their behaviors became more heavily driven by the market psychology, or their interpretation of the market psychology and reaction. Consequently, an algorithm that looks solely at the buy/sell imbalance sentiments becomes more effective. Though without more carefully designed experiments and quantitative data to back up this hypothesis, this observation only remains as an interesting hypothesis. If time permits, we plan to assess how our model behaves during different market environments.

Note a potentially alternate form of this model defines R(t) as follows:

(\*2)

The differences between (\*) and (\*2) includes:

1) If we are a market maker and want to use these models to give new quote mid level, (\*) produces price changes as time passed by since last quote. The longer time passed since last quote, the similar our new quote mid will be to the last quote mid. (\*2) produces a single result for the new quote mid that is irrelevant to the time passed since last quote.

2) The log transformation changes the linear relationship assumed in the regressions. This will potentially change the regression R^2, residuals pattern and etc.

Lastly, we want to assess the effectiveness of buy/sell/neutral classification rule #3:

We presented the results on AGN and CAT in the following charts. The results are very similar to the former method, though majority of them are lower than those of the former. It is very interesting that for AGN, the former method produces poor prediction result (best adj R^2: 1.3%) on 29th while this EMA quotes classification produces very good result (best adj R^2: 31.8%); for CAT, results of this new classification method on 25th and 26th are significantly higher than the former one.

Thus, we suggest we can further explore these two methods. Assess the parameter stability and model robustness.

AGN

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **4/23/13** | | **4/24/13** | | **4/25/13** | | **4/26/13** | | **4/29/13** | |
| *Concurrent Regression* | | | | | | | | | |
| Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 |
| 0.03\_1.25 | 48.6% | 0.1\_1 | 9.0% | 0.01\_0.75 | 6.2% | 0\_1 | 12.0% | .005\_1.25 | 9.6% |
| 0.03\_1.5 | 48.6% | 0.1\_0.75 | 9.0% | 0.01\_0.5 | 6.1% | 0\_1.25 | 12.0% | 0.005\_1.5 | 9.6% |
| 0.03\_1 | 48.6% | 0.1\_1.25 | 9.0% | 0.01\_1.25 | 6.1% | 0\_1.5 | 12.0% | 0\_0.005\_1 | 9.5% |
| 0.03\_0.75 | 48.6% | 0.1\_1.5 | 9.0% | 0.01\_1 | 6.1% | 0\_0.75 | 11.9% | 0.005\_0.5 | 9.5% |
| 0.03\_0.5 | 48.4% | 0.1\_0.5 | 8.5% | 0.01\_1.5 | 6.0% | 0\_0.5 | 11.9% | .005\_0.75 | 9.5% |
| *One Step Ahead Prediction* | | | | | | | | | |
| Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 |
| 0.03\_1.25 | 32.2% | 0.005\_1.5 | 4.2% | 0.025\_0.5 | 1.9% | 00\_0.08\_1 | 2.1% | 0\_1 | 31.8% |
| 0.03\_1.5 | 32.2% | 0.005\_1 | 4.1% | .025\_0.75 | 1.8% | 0.08\_1.25 | 2.1% | 0\_0.75 | 31.8% |
| 0.03\_1 | 32.2% | .005\_1.25 | 4.1% | 0.025\_1 | 1.8% | 0.08\_1.5 | 2.1% | 0\_0.5 | 31.8% |
| 0.03\_0.75 | 32.2% | .005\_0.75 | 4.1% | 0.025\_1.5 | 1.8% | 0.08\_0.5 | 2.1% | 0\_1.5 | 31.8% |
| 0.03\_0.5 | 32.2% | 0.005\_0.5 | 4.1% | .025\_1.25 | 1.8% | 0.08\_0.75 | 2.1% | 0\_1.25 | 31.8% |

CAT

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **4/23/13** |  | **4/24/13** |  | **4/25/13** |  | **4/26/13** |  | **4/29/13** |  |
| *Concurrent Regression* | | | | | | *Concurrent Regression* | | | |
| Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 |
| 0.015\_0.5 | 9.8% | 0.05\_1.5 | 10.8% | 0.055\_1 | 44.7% | 0.04\_0.75 | 42.2% | 0.045\_1.5 | 7.8% |
| 0.015\_1 | 9.7% | 0.01\_1.25 | 10.7% | 0.055\_1.5 | 44.7% | 0.04\_1 | 42.2% | .045\_1.25 | 7.6% |
| .015\_0.75 | 9.7% | 0.01\_1 | 10.7% | .055\_1.25 | 44.7% | 0.04\_1.25 | 42.1% | 0.04\_1.5 | 7.5% |
| .015\_1.25 | 9.5% | 0.01\_1.5 | 10.7% | .055\_0.75 | 44.7% | 0.04\_1.5 | 42.1% | .045\_0.75 | 7.3% |
| 0.015\_1.5 | 9.5% | 0.01\_0.75 | 10.6% | 0.055\_0.5 | 44.7% | 0.04\_0.5 | 41.9% | 0.045\_1 | 7.3% |
| *One Step Ahead Prediction* | | | | | | *One Step Ahead Prediction* | | | |
| Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 |
| .015\_1.25 | 9.1% | 0.03\_1 | 11.3% | 0.055\_0.5 | 42.9% | 0.04\_1.25 | 33.3% | 0.05\_0.5 | 6.3% |
| 0.015\_1 | 9.1% | 0.03\_1.25 | 11.2% | 0.055\_1 | 42.7% | 0.04\_1.5 | 33.3% | 0.05\_1.5 | 6.1% |
| 0.03\_0.5 | 9.1% | 0.03\_1.5 | 11.1% | .055\_1.25 | 42.7% | 0.04\_1 | 33.3% | 0.05\_1.25 | 6.1% |
| 0.015\_1.5 | 9.1% | 0.03\_0.5 | 11.1% | 0.055\_1.5 | 42.7% | 0.04\_0.5 | 33.3% | 0.05\_0.75 | 6.0% |
| 0.03\_1 | 9.0% | 0.03\_0.75 | 11.1% | .055\_0.75 | 42.7% | 0.04\_0.75 | 33.3% | 0.05\_1 | 6.0% |

**To do:**

The limited processing power stagnated the expansion of analysis to more stocks. We can only expand our analysis to small volume (~4 million/day) stocks so far. For larger volume stock like C (~20 million/day), MAC cannot change memory limit so Jia cannot parse those stocks; and it is also not feasible on Zhenyu’s laptop because it takes too long to parse. The solutions including: 1) Resample the data, only take part 1/10. 2) Use new data class to deal with big matric. (Packages like: bigmemory, biganalytics, synchronicity, bigalgebra, biglm and bigtabulate) This may calls for changing all the functions to handle the new type of inputs/outputs.

## Appendix I

**Formula (\*) regression results**

BA results:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **4/23/13** | | **4/24/13** | | **4/25/13** | | **4/26/13** | | **4/29/13** | |
| *Concurrent Regression* | | | | | | | | | |
| Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 |
| 0.005\_0.5 | 8.1% | 0.005\_0.5 | 10.6% | 0\_0.5 | 8.5% | 0.04\_0.5 | 3.9% | 0.01\_0.5 | 8.7% |
| 0.01\_1.5 | 8.1% | 005\_0.75 | 10.3% | 0.005\_0.5 | 4.6% | 0.04\_0.75 | 3.8% | 0.015\_0.5 | 8.0% |
| 0.01\_1.25 | 7.8% | 0.005\_1 | 8.7% | .005\_0.75 | 4.5% | 0.01\_0.75 | 3.8% | 0.005\_0.5 | 7.8% |
| 0.025\_0.5 | 7.5% | 0.02\_0.75 | 8.6% | 0.005\_1.5 | 4.4% | 0.01\_1 | 3.5% | 0.025\_0.5 | 7.2% |
| 0.01\_1 | 7.5% | 0.015\_0.75 | 8.5% | 0.005\_1 | 4.4% | 0.01\_0.5 | 3.5% | 0.01\_0.75 | 7.2% |
| *One Step Ahead Prediction* | | | | | | | | | |
| Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 |
| 0.02\_0.5 | 5.8% | 0.02\_0.75 | 13.4% | 0\_0.5 | 10.3% | 0.005\_1 | 2.0% | 0.01\_0.5 | 9.3% |
| 0\_1.5 | 4.9% | 0.02\_0.5 | 13.1% | 0.04\_0.5 | 6.6% | 0.005\_1.25 | 2.0% | 0.005\_0.5 | 9.1% |
| 0\_1.25 | 4.8% | 0.01\_0.75 | 13.0% | 0.095\_0.5 | 6.5% | 0.005\_0.75 | 1.9% | 0.015\_0.5 | 8.7% |
| 0\_1 | 4.6% | 0.03\_0.75 | 12.8% | .095\_0.75 | 6.4% | 0.005\_1.5 | 1.9% | 0.045\_0.5 | 7.1% |
| 0.055\_0.5 | 4.5% | 0.005\_0.75 | 12.7% | 0.07\_0.5 | 6.3% | 0.005\_0.5 | 1.5% | 0.025\_0.5 | 6.9% |

CAT:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **4/23/13** | | **4/24/13** | | **4/25/13** | | **4/26/13** | | **4/29/13** | |
| *Concurrent Regression* | | | | | | | | | |
| Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 |
| 0.09\_0.5 | 10.2% | 0.05\_1.25 | 13.0% | 0.025\_1.5 | 13.7% | 0.095\_0.5 | 10.8% | 0.04\_0.5 | 7.9% |
| 0.085\_0.5 | 9.5% | 00\_0.05\_1 | 13.0% | 0.05\_1.5 | 13.1% | 0.09\_0.5 | 10.7% | 0.045\_0.5 | 7.8% |
| 0.095\_0.5 | 9.1% | 0.05\_1.5 | 13.0% | 0.045\_1.5 | 12.6% | .095\_0.75 | 9.6% | 0.04\_0.75 | 7.8% |
| 0.1\_0.5 | 9.0% | 0.05\_0.75 | 12.9% | .025\_1.25 | 12.4% | 0.09\_0.75 | 9.3% | 0.055\_0.5 | 7.7% |
| 0.09\_0.75 | 8.6% | 0.05\_0.5 | 6.1% | 0.055\_1.5 | 12.4% | 0.095\_1 | 8.7% | .045\_0.75 | 7.7% |
| *One Step Ahead Prediction* | | | | | | | | | |
| Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 |
| 0\_0.5 | 73.1% | 0.01\_0.5 | 7.1% | 0.025\_1.5 | 32.6% | 0.09\_0.5 | 6.4% | 0.04\_0.5 | 7.5% |
| 0\_0.75 | 62.5% | 0.01\_0.75 | 6.2% | 0.05\_1.5 | 31.5% | 0.095\_0.5 | 6.4% | 0.045\_0.5 | 7.3% |
| 0\_1 | 55.0% | 0.03\_0.75 | 5.6% | 0.03\_1.5 | 31.5% | 0.07\_0.5 | 5.3% | 0.04\_0.75 | 7.0% |
| 0\_1.25 | 42.6% | 0.03\_0.5 | 5.6% | .025\_1.25 | 31.0% | 0.045\_0.5 | 5.2% | 0.055\_0.5 | 7.0% |
| 0.01\_0.5 | 36.1% | 0.01\_1 | 5.5% | 0.055\_1.5 | 30.5% | 0.05\_0.5 | 5.0% | .045\_0.75 | 6.8% |

TIF

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **4/23/13** | | **4/24/13** | | **4/25/13** | | **4/26/13** | | **4/29/13** | |
| *Concurrent Regression* | | | | | | | | | |
| Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 |
| 0.065\_1 | 8.1% | 0.06\_0.5 | 21.0% | 0.075\_1.5 | 5.8% | 0.075\_1.5 | 12.2% | 0\_0.75 | 33.9% |
| 0\_0.04\_1 | 8.1% | 0.065\_0.5 | 20.9% | 0.08\_1.5 | 5.8% | .075\_1.25 | 12.2% | 0\_0.5 | 33.7% |
| 0\_0.03\_1 | 8.1% | 0.095\_0.5 | 20.8% | 0.07\_1.5 | 5.8% | 0.085\_1.5 | 12.0% | 0\_1 | 31.6% |
| 0.045\_1 | 8.1% | 0.1\_0.5 | 20.8% | .075\_1.25 | 5.7% | .085\_1.25 | 12.0% | 0.09\_0.5 | 29.4% |
| 0.035\_1 | 8.0% | 0.075\_0.5 | 20.7% | 0.085\_1.5 | 5.7% | 0.08\_1.5 | 11.9% | .095\_0.5 | 29.3% |
| *One Step Ahead Prediction* | | | | | | | | | |
| Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 | Key | Adj R2 |
| 0.02\_0.5 | 5.2% | 0.015\_1.5 | 10.9% | 0.095\_1.5 | 5.2% | 0.055\_1.5 | 11.3% | 0\_0.5 | 10.1% |
| 0.025\_0.5 | 5.0% | .015\_1.25 | 10.7% | 0.1\_1.5 | 5.2% | 0.015\_1.5 | 11.0% | 0.1\_0.5 | 10.1% |
| 0.01\_0.5 | 4.8% | 0.02\_1 | 10.5% | 0.09\_1.5 | 5.2% | 0.045\_1.5 | 10.5% | 0.08\_0.5 | 10.0% |
| 0.015\_0.5 | 4.6% | 0.02\_1.25 | 10.5% | .095\_1.25 | 5.1% | .055\_1.25 | 8.5% | 0.085\_0.5 | 9.9% |
| 0.065\_0.5 | 4.2% | 0.02\_1.5 | 10.2% | 0.1\_1.25 | 5.1% | .015\_1.25 | 8.3% | 0.095\_0.5 | 9.8% |

## Second Part: where stocks trade?

In the second part of the project, we are trying to understand the factors affecting the distribution of liquidity across trading venues. In modern US equity market, securities can be traded on multiple venues and each trading venue has different liquidity characters. Since the market depth of each trading venue has great influence on order execution quality, this study has important implications for optimal order placement. The target of the study is to construct a robust model to predict the optimal number of shares to be placed across trading venues with potential input of security symbol and number of shares to be bought or sold. Table one is the list for current available trading venues in the US.

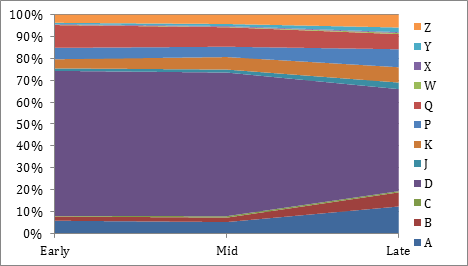
|  |  |  |  |
| --- | --- | --- | --- |
| **Symbol** | **Trading Venue** | **Symbol** | **Trading Venue** |
| A | NYSE MKT | N | NYSE Euronext |
| B | NASDAQ OMX BX | P | NYSE Arca Exchange |
| C | National Stock Exchange | Q | NASDAQ OMX |
| D | FINRA Trade Reporting Facility | W | Chicago Board Options Exchange |
| J | EDGA Exchange | X | NASDAQ OMX PHLX |
| K | EDGX Exchange | Y | BATS Y-Exchange |
| M | Chicago Stock Exchange | Z | BATS Exchange Inc |

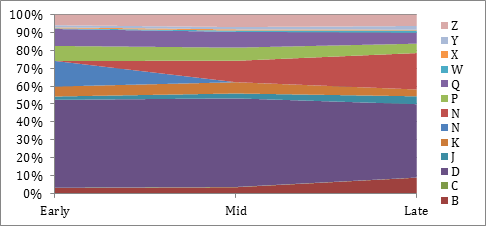
*Table 1: Symbols for Trading Venues*

Given there are more than 7000 stocks listed across all trading venues, it will be unrealistic to analyze the problem starting from all 7000 stocks, we decide to perform analysis for AAPL on 20130424 trading data which has significant volume across major trading venues. Once meaningful results is achieved for AAPL, we will expand our analysis to all 7000 stocks and multiple trading days. Also we noticed that the trading behavior during opening auction and closing auction is different from the trading behavior during other market hours, thus we divided the trading day into three sub groups: EDT 9:30:00 - 9:39:59 (UTC 13:30:00 - 13:39:59), EDT 9:40:00 - 15:49:59 (UTC 13:40:00 - 19:49:59), EDT 15:50:00 - 16:00:00 (UTC 19:50:00 - 20:00:00) corresponds to “early” time group, “midday” time group and “late” time group.

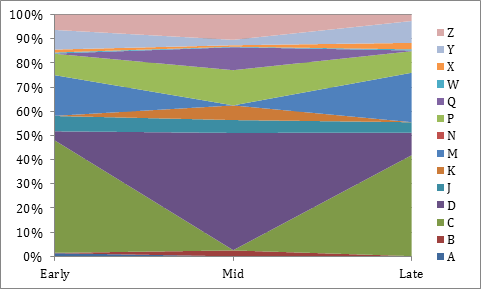
In the first step of the analysis, we are interested to see the trading volume distribution across trading venues for the three time groups.

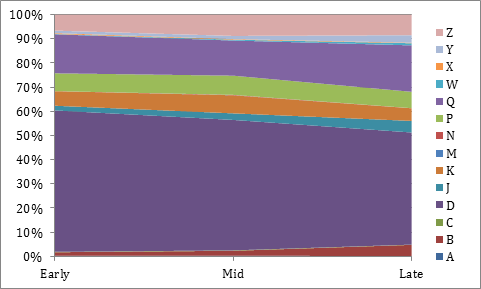
First we aggregated all security symbols according to primary trading venue, we treat all securities have the same primary trading venue as a group. We analyzed trade volume across trading venues and presented in the following charts:

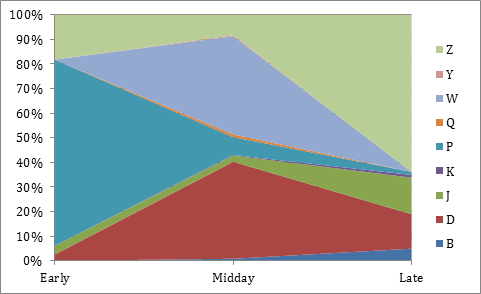
Primary trading venue A group:

Primary trading venue N group:

Primary trading venue P group:



Primary trading venue Q group:

Primary trading venue Z group:

With the exception of the Z primary listed securities (primarily ETFs), we can observe D dominating midday and early and late period exhibits similar behavior.  Second most traded venue tends to be the primary trading venue.

Furthermore, we decompose the trade volume within D group to D1 and D2, D1 is where trade price occurs at midpoint of bid/ask price with tolerance of 1 cent, D2 are other traders happened in D. The rationale is to separate dark pool flows from the rest of D since the quotes are not shown in these market, the trades tend to occur outside of the mid-price range.

Decomposition of AAPL:

Venue                     Early          Midday         Late

D1(midpoint)          0.02%        1.46%          0.78%

D2(nonmid point)   54.35%      54.81%        55.81%

Q(primary venue)   9.88%       11.01%        16.08%

others                     35.75%      32.71%        27.33%

Since AAPL has a relatively high stock price compared to other securities, we are afraid that the result may be biased, so we conducted analysis on BAC as well.

Venue                     Early          Midday         Late

D1(midpoint)          8.90%        16.69%        13.03%

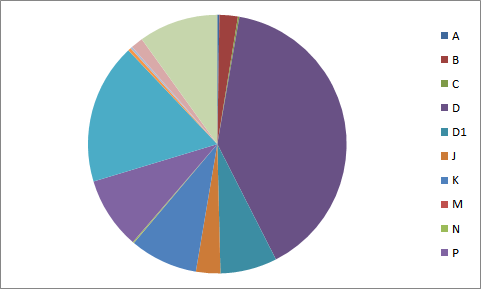
D2(nonmid point)   38.27%      41.41%        37.20%

N(primary venue)   4.81%        2.52%          4.31%

others                     48.02%      44.38%        45.56%

We can see the pattern still prevails, majority of D trades happens outside of the midpoint of bid/ask price.

**All stock analysis**



Plot of Exchange composition for tickers listed on Q

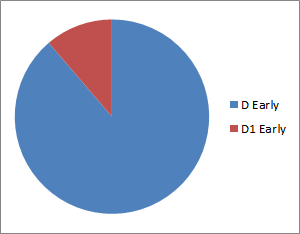
Here is an analysis for the all the stocks primary listed on Exchange Q.  We can see that the trend of the aggregate analysis follows the individual tickers shown before.  The D1 still makes up a relative small portion of all the D trade flows.  This could mean that the amount of off exchange activity made up a minority portion of all D flows.

**Other Factors**

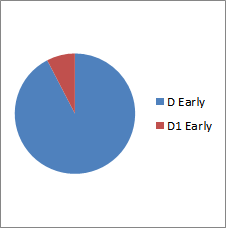
We also considered some other factors such as volatility, average price and company characteristics.

For each variable, for instance for the volatility, we take ten stocks with high volatility and ten with low volatility, aggregate them into two groups, and compare the trade flow composition.

High Vol



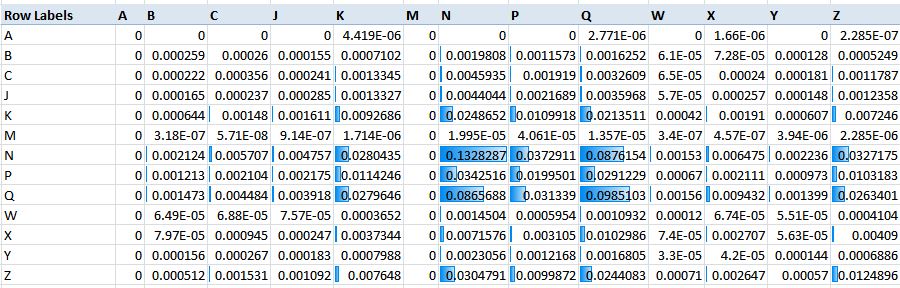
Low Vol



For the result, price/volatility factors don’t have an effect on the trading venue composition.  Their composition of trade flows looks quite similar.

**Best Quotes**

Plot of Best Bid/Ask frequency for all the tickers trades on exchange N, number normalized



The definition of best quotes are debatable.  We perform some analysis on the best bid/ask based on the last bid/ask that closest to the trade as a benchmark and also based on the amount of quotes distribution.  The result shows that N,P,Q are the most popular quote exchange by best bid/ask and by amount of quotes.  Beside the N/P/Q, the amount of quotes we observe are distributed quite evenly among other remaining exchanges.

**Contingency analysis**

Contingency table is a table in matrix format that displays the frequency distribution of variables.

Chi-square test for any contingency table can be written as

where Oi is observation, and Ei=

It can be used to test the independence null hypothesis. One problem with chi-square test is that large sample size will inflate chi-square statistic, and always rejects the independence null hypothesis.

Cramér's V is another popular measure of association between two nominal variables, giving a value between 0 and 1.

Cramér's V = where N is the total number of observations and k is the minimum of row number and column number. Interpretation of Cramér's V is like correlation.

|  |  |
| --- | --- |
| **Cramér's V** | **Interpretation** |
| [0, 0.05] | No relationship |
| (0.05, 0.1] | Weak relationship |
| (0.1, 0.15] | Moderate relationship |
| (0.15, 0.25] | Strong relationship |
| (0.25, 1] | Very strong relationship |

*Table 1: Interpretation of Cramér's V*

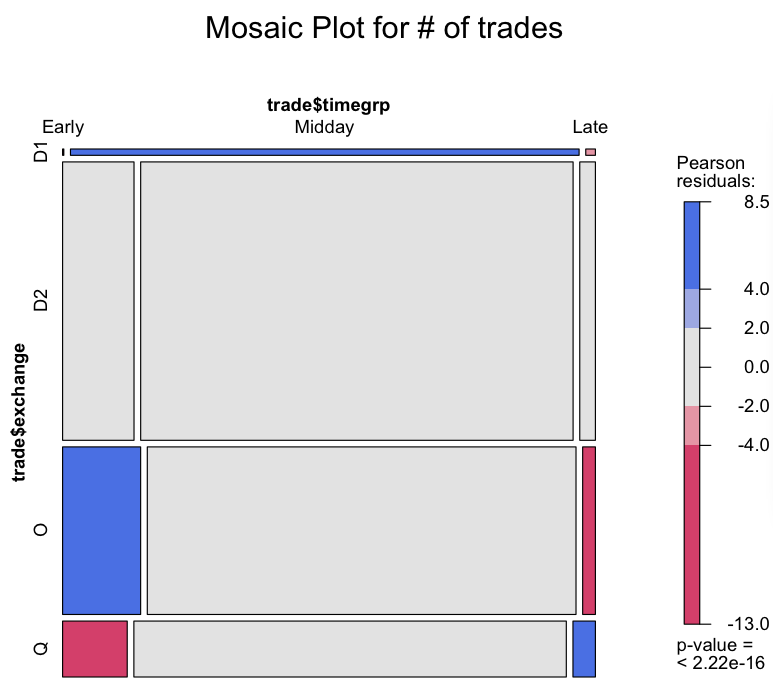
The raw data for contingency analysis is the transaction records for AAPL on April 24th. We focus on studying the relationship between when the security traded and where the security traded.

* We divide the transaction time into three categories: Early (take place in the first 10 minutes), Late (take place in the last 10 minutes), and Midday (other time).
* We divide the trading venue into four categories: D1 (non block-trade in darkpool), D2 (block-trade in darkpool), Q (NASDAQ, which is the primary exchange of AAPL), O (all the other trading venues)

Contingency table for number of transactions for AAPL is displayed in *Table 2*, and mosaic plot of this contingency table is displayed in *Figure 1*.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Early | Midday | Late |
| D1 | 3 | 1265 | 24 |
| D2 | 7828 | 47400 | 1715 |
| O | 5149 | 28287 | 840 |
| Q | 1423 | 9523 | 494 |
| Cramer’s v = 0.04832802 – No Relationship | | | |

*Table 2: Contingency Analysis for Number of Transaction*



*Figure 1: Mosaic Plot for Number of transactions*

Interpretation of mosaic plot is straightforward. We interpret positive Pearson residuals (label in blue) as showing cells whose observed frequency is substantially greater than would be found under independence; negative Pearson residuals values (label in red) indicate cells which occur less often than under independence.

In *Figure 1*, most of the cells are not significant. We should pay attention to D1 in midday. About 1265 D1 transactions take place in midday, which is around 97.9% of the total D1 transactions.

In *Table 3*, we construct a contingency table in term of number of shares instead of number of transactions.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Early | Midday | Late |
| D1 | 771 | 289821 | 5135 |
| D2 | 1482682 | 11042741 | 368502 |
| O | 633770 | 3818988 | 98148 |
| Q | 180865 | 1411640 | 83055 |
| Cramer’s v = 0.04946721 – No Relationship | | | |

*Table 3: Contingency Analysis for Number of Shares*

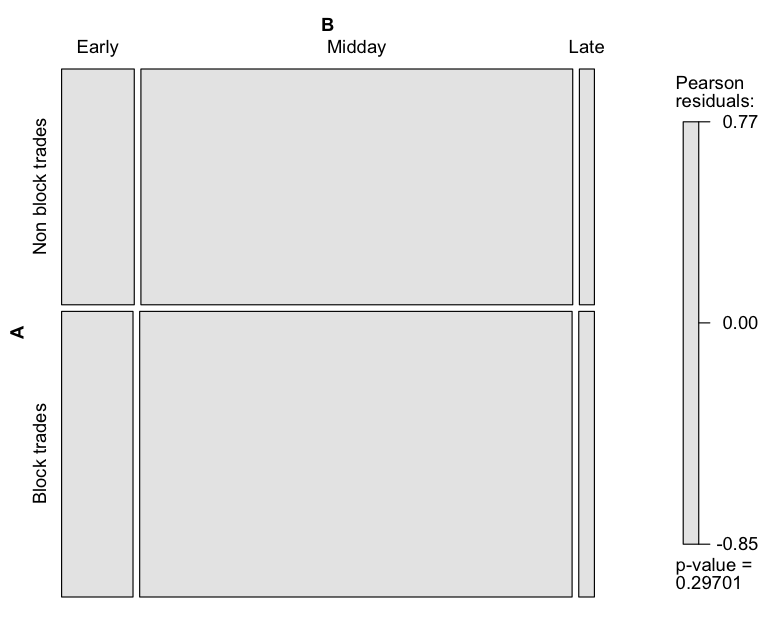
The result of *Table 3* is similar to *Table 2*. Values of Cramer’s v are close, and the number of shares trade in darkpool as non-block (D1) is unproportionate large in midday.

There are some other venue/time nodes that are significant according to the mosaic plot. However, the significance is due to large number of transactions.

Comparing block trade with all the other trades (combine D1, Q, and O), the results are displayed in *Table 4*, *Figure 2*, and *Table5*.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Early | Midday | Late |
| Non-D2 | 6575 | 39075 | 1358 |
| D2 | 7828 | 47400 | 1715 |
| Cramer’s v = 0.004832872 – No Relationship | | | |

*Table 4: Contingency Analysis for Number of Transaction*



*Figure 2: Mosaic Plot for Number of transactions*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Early | Midday | Late |
| Non-D2 | 815406 | 5520449 | 186338 |
| D2 | 1423 | 9523 | 494 |
| Cramer’s v = 0.0507463 – Weak Relationship | | | |

*Table 5: Contingency Analysis for Number of Shares*

Values of Cramer’s v are around 0.05, and all the blocks in mosaic plot are insignificant. We can conclude that trading venues are independent of when the security trade.

**K-means Clustering**

* **Methodology**

K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

We are given observations X1,…,Xn, and dissimilarites d(Xi;Xj) for i; j = 1,…,n.

d(Xi;Xj) = (Euclidean space)

Let K denote the number of clusters, and dij = d(Xi,Xj). The within-cluster scatter is defined as

The within-cluster scatter can be rewritten as

with is the average of points in group k,

We seek a clustering C minimizing the within-cluster variation.

The algorithm always converges, no matter what are the initial cluster centers. However the final result depends on the initial cluster centers. In some cases, different initial centers can lead to different final outputs. Hence it is a good idea to run K-means multiple times.

What we do here is trying to cluster all the securities in order to minimize the within-group variation (W(K)) and maximize between-group variation (B(K)). Securities in the same cluster behave similarly, so we can explore the characteristic of each group. Basically we use k-means clustering, and we would choose the clusters with highest CH value.

Therefore, CH value is an essential value to find the best k, and choose the result if return multiple clustering results.

* **Raw data**

For clustering analysis, we include all the securities listed on A, N, P, Q, and Z. Each row represents one security with a unique ticker, and each column represents a trading venue. The cell value is the trading volume of the certain security took place in the certain trading venue on April 24th 2013, and we treat the first ten minutes (early), last ten minutes (late), and other time period (midday) separately.

We then standardize the data by dividing each element by row total, so the sum of each row equals to 1. Each cell represents the proportion of trading volume of the certain security trades in the certain venue to the security’s total trading volume in one day. The sample raw data is displayed in Table 7.

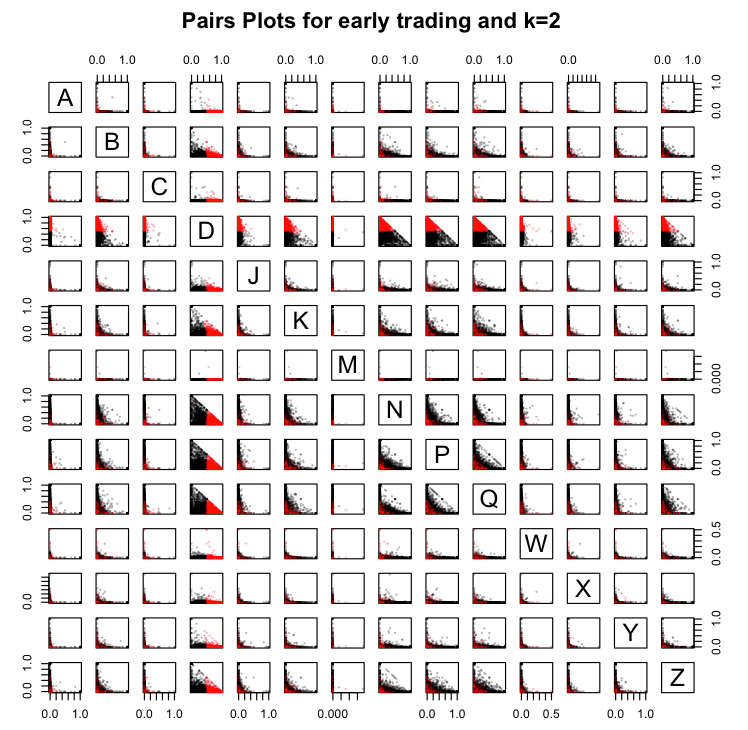
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Tickers | A | B | C | D | J | K | M | … |
| A | 0 | 0.0218 | 0.0020 | 0.1875 | 0.0060 | 0.0257 | 0 | … |
| AA | 0 | 0.0230 | 0.0087 | 0.6021 | 0.0297 | 0.0358 | 0 | … |
| … | … | … | … | … | … | … | … | … |

*Table 7: Sample Raw Data*

* **Result and Interpretation**

For all the three time periods, CH value rises first and then starts to decrease after K=6. So we will mainly study K=2,3,4,5,6.

For convenience, each cluster will be given a name based on its feature. For example, if all the securities in one cluster traded more in X, we will call this cluster “Cluster X”. If there is no specific feature in the cluster, we will call it “Cluster O”



*Figure 3: Clustering plot of early trading and k=2*

The determent factor in K=2 k-means clustering is dimension D. If most of the trades of a certain security happen in D, it might be labeled red; if most of the trades are done in other venues, the security will be labeled black in the plot. The basic information for each cluster is summarized in *Table 8-9*.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster Name** | **Color** | **# of securities** | **Index of cluster center** | | | | | | |
| O | Black | 2196 | A | B | C | D | J | K | M |
| 0.008 | 0.040 | 0.003 | 0.214 | 0.018 | 0.062 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.190 | 0.179 | 0.176 | 0.003 | 0.004 | 0.015 | 0.088 |
| D | Red | 2305 | A | B | C | D | J | K | M |
| 0.001 | 0.015 | 0.002 | 0.762 | 0.008 | 0.027 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.039 | 0.044 | 0.062 | 0.002 | 0.002 | 0.006 | 0.029 |

*Table 8: Cluster Centers for Early Trading and K=2*

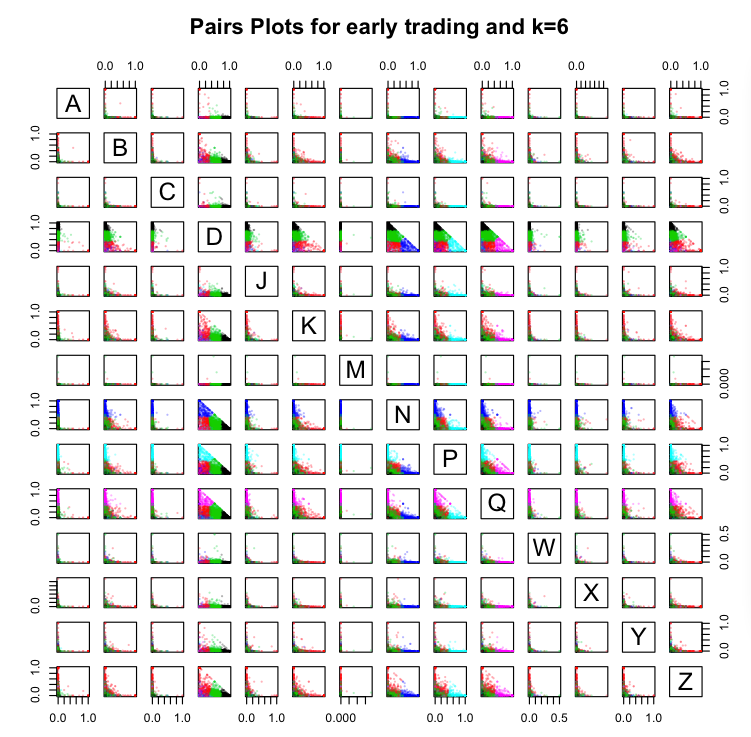
* Numbers of securities in two clusters don’t vary much.
* For the centers of the two clusters, the difference of values on dimension D is significant, which is consistent with the observation in *Figure 3*
* The mid point in other dimensions are really small

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster Name** | **Color** | **Top 5 securities based on total trading volume** | **Statistics for Total Trading Volume** | | |
| O | Black | MSFT, XLF, EEM, PFE, F | Max | Min | Average |
| 1941695 | 100 | 24292.12 |
| D | Red | BAC, SIRI, UVXY, CLSN, COH | Max | Min | Average |
| 9770460 | 100 | 40184.48 |

*Table 9: Trading Volume of Clusters for Early Trading and K=2*

* Generally speaking, the trading volume for Cluster O is less than that of Cluster D. However, if we do a t test to test the equality of means of two clusters, t statistic is -2.944 and p-value is 0.003265. Using 0.05 as a threshold, the means of total trading volumes of two clusters are significantly different.

Then we change the value of k for early trading. As k increases, there more and more factors become significant. When K=6, D, N, P, Q are significant factors according to the pair plots. In dimension D, there are two cutoffs.



*Figure 4: Clustering plot of early trading and k=6*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster Name** | **Color** | **# of securities** | **Statistics for Trading Volume** | | | | | | |
| D1 | Black | 1198 | A | B | C | D | J | K | M |
| 0.000 | 0.005 | 0.000 | 0.914 | 0.003 | 0.010 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.014 | 0.018 | 0.021 | 0.000 | 0.000 | 0.003 | 0.010 |
| O | Red | 717 | A | B | C | D | J | K | M |
| 0.000 | 0.008 | 0.000 | 0.171 | 0.027 | 0.125 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.104 | 0.114 | 0.134 | 0.004 | 0.006 | 0.026 | 0.183 |
| D2 | Green | 1714 | A | B | C | D | J | K | M |
| 0.000 | 0.029 | 0.000 | 0.532 | 0.016 | 0.047 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.092 | 0.090 | 0.117 | 0.004 | 0.004 | 0.012 | 0.053 |
| N | Blue | 336 | A | B | C | D | J | K | M |
| 0.000 | 0.020 | 0.000 | 0.110 | 0.007 | 0.018 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.740 | 0.034 | 0.035 | 0.003 | 0.002 | 0.006 | 0.022 |
| P | Light Blue | 277 | A | B | C | D | J | K | M |
| 0.000 | 0.008 | 0.000 | 0.107 | 0.011 | 0.014 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.020 | 0.769 | 0.031 | 0.001 | 0.001 | 0.007 | 0.031 |
| Q | Purple | 259 | A | B | C | D | J | K | M |
| 0.000 | 0.014 | 0.000 | 0.110 | 0.006 | 0.028 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.008 | 0.056 | 0.729 | 0.001 | 0.001 | 0.008 | 0.038 |

*Table 10: Cluster Centers for Early Trading and K=6*

* Most of securities are in Cluster D1(26.6%), O(15.9%) and D2(38.1%). These 3 clusters are divided mainly based on dimension D.
* The center points for each cluster are consistent with observation from the pair plots. For example, securities in Cluster D1(Black) seem to be mostly traded in D, and the center of this cluster has high value on dimension D. The significant values in each cluster have been labeled red.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster Name** | **Color** | **Top 5 securities based on total trading volume** | **Statistics for Total Trading Volume** | | |
| D1 | Black | SIRI, CLSN, LCC, NXPI, SHFL | Max | Min | Average |
| 1975365 | 100 | 16421.27 |
| O | Red | XLF, PFE, EWJ, RF, CSCO | Max | Min | Average |
| 1726286 | 100 | 26136.36 |
| D2 | Green | BAC, MSFT, UVXY, COH, NFLX | Max | Min | Average |
| 9770460 | 200 | 59011.52 |
| N | Blue | MRK, PG, DOW, WAG, FBR | Max | Min | Average |
| 785947 | 100 | 8698.94 |
| P | Light-Blue | FXI, EWT, EWS, DXJ, EWH | Max | Min | Average |
| 373773 | 100 | 8883.74 |
| Q | Purple | GLUU, CLMT, JDSU, PPC, EXEL | Max | Min | Average |
| 119009 | 100 | 3972.521 |

*Table 11: Trading Volume of Clusters for Early Trading and K=6*

* Generally speaking, the trading volume for Cluster D1, O and D2 are greater than that of cluster N, P and Q. Cluster D2 (Green) has the highest trading volume, and Cluster Q (Purple) has the lowest trading volume.
* Some t-tests to test the equality of mean total trading volume.

1. Cluster D2 vs other clusters: t = 6.2166, p-value = 6.235e-10
2. Cluster D1 vs Cluster O: t = -2.142, p-value = 0.03237
3. Cluster N vs Cluster P: t = -0.0568, p-value = 0.9547
4. Cluster Q vs other clusters: t = -10.0882, p-value < 2.2e-16

From the t-tests above, we can order the clusters according to average total trading volume: Cluster D2 Cluster D1 & O Cluster N & P Cluster Q.

Then we apply the same method to midday and late transactions as well. As K increases, there are more significant factors. For the pair plots for each K in each time period, please see *Figure* in the attachment.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Early | Midday | Late |
| 2 | D1 | D1 | D |
| 3 | D1,N | D1,P | D,N |
| 4 | D1,N,P | D1,P,D2 | D,N,Q |
| 5 | D1,N,P,Q | D1,P,D2,N,Q | D,N,Q,P |
| 6 | D1,N,P,Q,D2 | D1,P,D2,N,Q,K | D,N,Q,P,K |

*Table 12: Significant Factors According to Clustering Pair Plots*

Result of K=2 for midday and late are summarized inthe . There are several observations:

* D is also a key factor in midday and late clustering.
* Average total trading volume of Cluster O (less trade in D) is greater than that of Cluster D (more trade in D) in midday, and in last ten minute are close to each other (We cannot reject equality hypothesis under =0.01). These results are not consistent with early trading. One possible explanation is that market markers are more active in the first 10 minutes than the rest of the day; institutional investors and retail investors have less trading activities in this time period.
* EEM is always among the top five securities in Cluster O based on trading volume.
* BAC has gargantuan trading volume through the whole day. Although its cluster changes, the percentage of trading volume in D remains consistent (early: 48.64%, midday: 57.27%, late: 50.47%)

Result of K=6 for midday and late are summarized in the attachment. There are several observations:

* 6 clusters in all three time-periods share the same features. Dimension D, P, Q, N play efficient roles in clustering. In last 10 minutes, trading venue K becomes significant as well. Here we specify the names for each clusters:

Midday: Cluster D1, Cluster D2, Cluster N, Cluster P, Cluster Q, Cluster O

Late: Cluster D1, Cluster D2, Cluster N, Cluster P, Cluster Q, Cluster K

* We order the clusters according to average total trading volume:

Early: Cluster D2 Cluster D1 & O Cluster N & P Cluster Q

Midday: Cluster D2 Cluster N & O Cluster D1 & P Cluster Q

Late: Cluster D2 Cluster N & P Cluster D1 & Q Cluster K

Cluster D2 always has the highest average total trading volume. Trading volumes of the securities in Cluster D1 become less in midday and late time compared with other clusters.

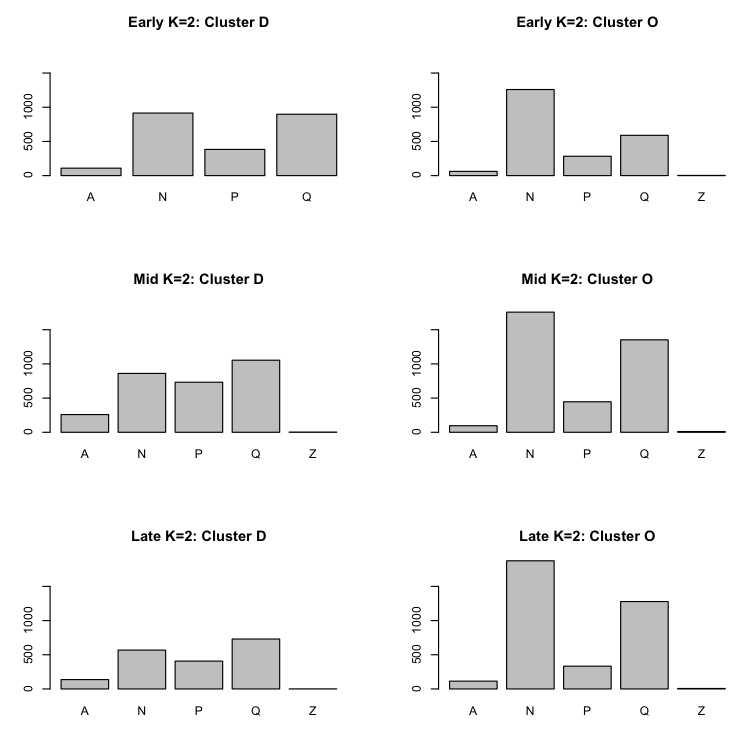
* For all three time periods, most of the securities are in Cluster D1 and D2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Number of securities in cluster D1 | Percentage | Number of securities in cluster D2 | Percentage |
| Early | 1198 | 26.6% | 1714 | 38.1% |
| Midday | 1397 | 21.3% | 2180 | 33.2% |
| Late | 1155 | 21.2% | 2293 | 42.1% |

*Table 13: Number of Securities in Cluster D1 and D2*

* **Compare Securities’ Primary Trading Venue through Clusters**

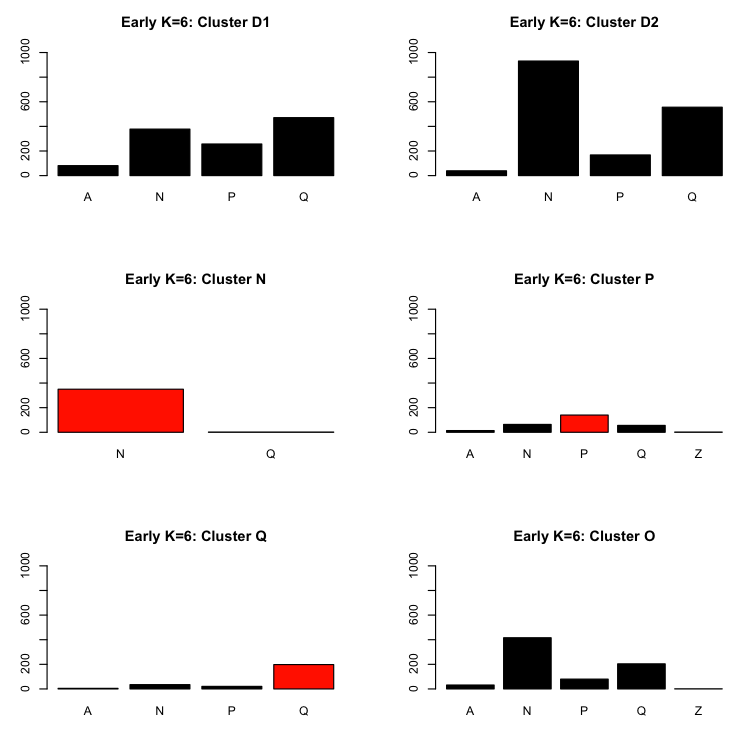
This part will display the distribution of securities’ primary exchange in each cluster.



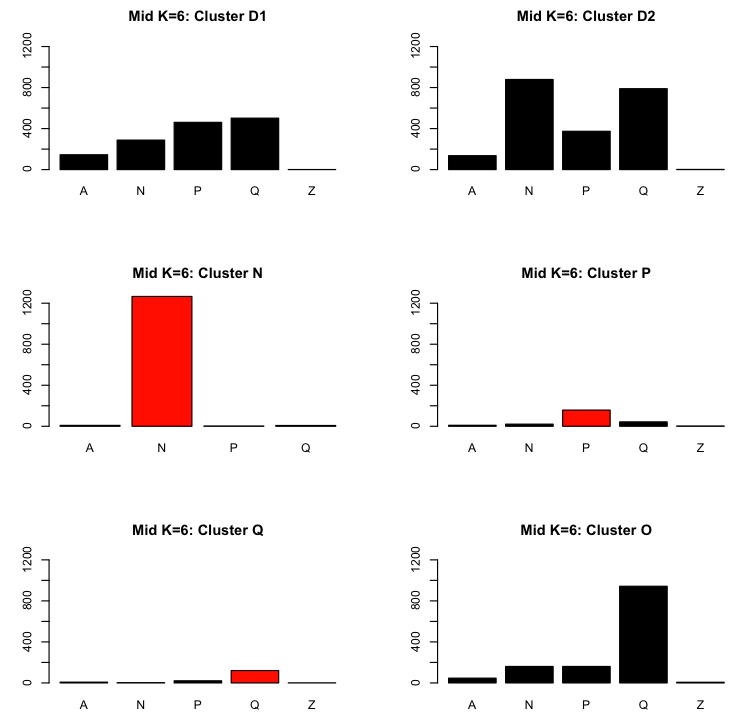
*Figure 5: Barplots of Securities’ Primary Exchanges in All Clusters*

*Figure 5* is a matrix plot, and display primary exchange distribution for K=2 k-means clustering in early, mid and late.

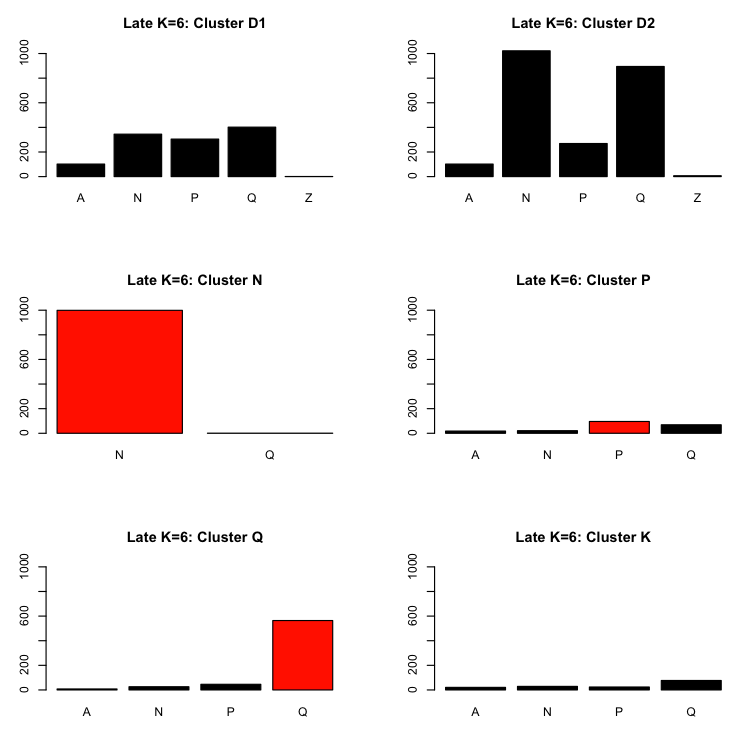
For securities listed on A and P, most of them are in Cluster D rather than Cluster O. However, the difference is not significant. Most of the securities listed on N are in Cluster O.



*Figure 6: Barplots of Stocks’ Primary Exchange in 6 different clusters in early*



*Figure 7: Barplots of Stocks’ Primary Exchange in 6 different clusters in midday*



*Figure 8: Barplots of Stocks’ Primary Exchange in 6 different clusters in late*

*Figure 6 – 8* display primary exchange distribution for K=2 k-means clustering in early, mid and late.

* Most of the securities listed on A, P and Q are in Cluster D1 and D2
* Around half of securities listed on N are in Cluster D1 and D2; another half are in Cluster N.
* Almost all the stocks in Cluster N are listed on N.
* Most of the stocks trades at Q are listed on Q.

**Principal Component Analysis**

Principal Component Analysis uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

The first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to the preceding components.

Therefore, one major character for these components is that they are uncorrelated with each other.

The raw data is the same as in the k-means clustering analysis. For early transactions, the first 8 principal components and the proportion of variance they can explain are summarized in *Table 14,* and the first 3 principal component loadings are displayed in *Table 15*.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **PC.1** | **PC.2** | **PC.3** | **PC.4** |
| **Standard deviation** | 1.285 | 1.11006 | 1.06308 | 1.04310 |
| **Proportion of Variance** | 0.118 | 0.08802 | 0.08072 | 0.07772 |
| **Cumulative Proportion** | 0.118 | 0.20603 | 0.28676 | 0.36447 |
|  | **PC.5** | **PC.6** | **PC.7** | **PC.8** |
| **Standard deviation** | 1.01565 | 1.01360 | 1.00595 | 1.0040 |
| **Proportion of Variance** | 0.07368 | 0.07338 | 0.07228 | 0.0720 |
| **Cumulative Proportion** | 0.43816 | 0.51154 | 0.58382 | 0.6558 |

*Table 14: Deviation Principal Component can Explain for Early Trading Volume*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **PC1** | **PC2** | **PC3** |
| **A** | 0,03 | -0,023 | 0,068 |
| **B** | 0,181 | 0,191 | -0,307 |
| **C** | 0 | 0,011 | -0,034 |
| **D** | **-0,771** | -0,009 | -0,051 |
| **J** | 0,132 | -0,048 | -0,148 |
| **K** | 0,146 | -0,185 | -0,236 |
| **M** | 0,006 | -0,049 | -0,009 |
| **N** | 0,258 | **0,68** | 0,061 |
| **P** | 0,299 | -0,131 | **0,75** |
| **Q** | 0,29 | **-0,497** | **-0,352** |
| **W** | 0,072 | 0,272 | -0,194 |
| **X** | 0,097 | 0,27 | -0,173 |
| **Y** | 0,118 | 0,072 | -0,247 |
| **Z** | 0,256 | -0,211 | -0,037 |

*Table 15: First 3 Principal Component Loadings for Early*

The first 8 principal component directions can explain about two third of the variation in the data set. The proportions of variance each principal component can explain are very close.

In the first principal component, D is the dominant variable, and D is negatively correlated with all the other variables, since D is the only variable with negative sign; in the second principal component, N and Q become significant, and they are negatively correlated; in the third principal component, P becomes significant, and P is negatively correlated with all the other factors in this direction.

The observation from principal component analysis is similar with in clustering analysis. D, N, P, Q play important roles when explaining the variance of data.

When applying principal component analysis to midday and late data, the percentage of variance explained by each principal component direction is similar with in early analysis.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Midday** | | | **Late** | | |
| **PC1** | **PC2** | **PC3** | **PC1** | **PC2** | **PC3** |
| **A** | 0,03 | -0,096 | 0,023 | 0,025 | -0,074 | 0,149 |
| **B** | -0,352 | 0,305 | -0,084 | -0,37 | 0,323 | -0,25 |
| **C** | 0,004 | -0,103 | -0,015 | -0,082 | -0,064 | 0,087 |
| **D** | **0,584** | 0,27 | -0,26 | **0,701** | 0,29 | -0,21 |
| **J** | -0,28 | 0,098 | -0,205 | -0,166 | 0,032 | -0,334 |
| **K** | -0,172 | -0,232 | -0,17 | -0,048 | -0,186 | 0,185 |
| **M** | -0,002 | -0,018 | 0,063 | -0,005 | -0,056 | 0,012 |
| **N** | -0,231 | **0,493** | 0,31 | -0,309 | **0,438** | **0,513** |
| **P** | -0,135 | -0,303 | **0,74** | -0,156 | -0,328 | 0,374 |
| **Q** | -0,254 | **-0,472** | **-0,323** | -0,125 | **-0,567** | -0,254 |
| **W** | -0,101 | 0,089 | 0,068 | -0,275 | 0,257 | -0,189 |
| **X** | -0,268 | 0,389 | 0,012 | -0,1 | 0,136 | 0,075 |
| **Y** | -0,299 | 0,126 | -0,267 | -0,273 | 0,123 | **-0,399** |
| **Z** | -0,344 | -0,13 | -0,16 | -0,197 | -0,205 | -0,222 |

*Table 16: First 3 Principal Component Loadings for Midday and Late*

As for principal component analysis for midday and late, the results turn out to be similar as in the early. D, N, Q are significant. The only exception is in the third principal component direction of late period, where N and Y are the two significant factors in this direction, and they are negatively correlated.

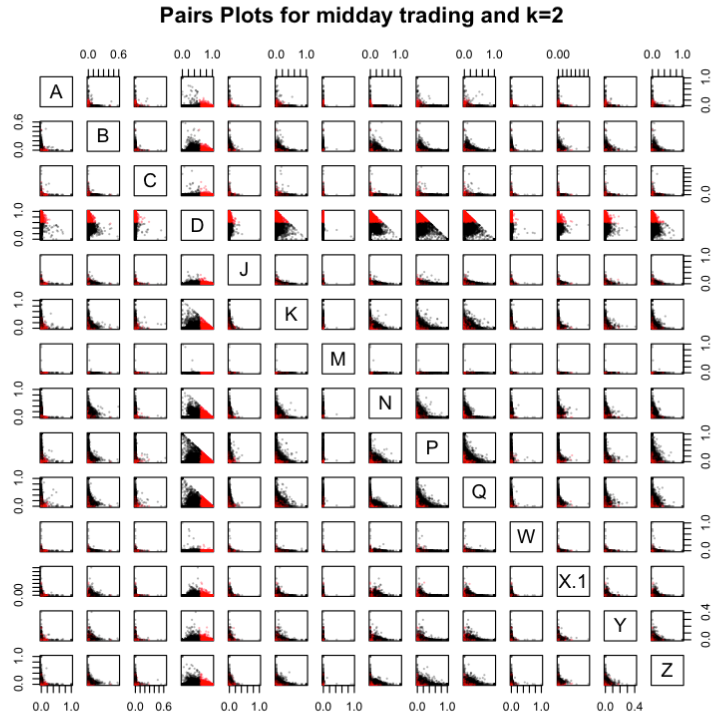
**Classification analysis**

This is the main part for the rest of this project. What we are doing in clustering analysis and principal component analysis give us intuition about the data, and classification analysis can provide us with a model, which can be used in real life.

What we will do in the next couple of weeks:

1. Select explanatory variables of the data; clean and process the data.
2. Apply different classification techniques to data, including multinomial logistic regression, decision trees, k-nearest neighbor, linear discriminant analysis, and etc.
3. Apply cross validation technique to each of the method, and compare the misclassification rate. We will finally choose the one with the lowest misclassification rate to be the final model.

## Attachment II



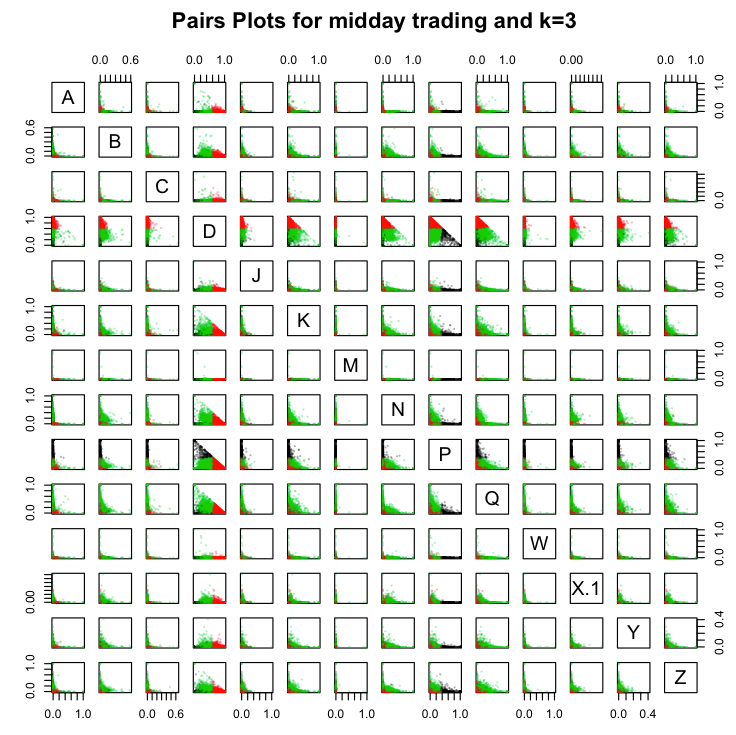
*Figure 9: Clustering plot of midday trading and k=2*

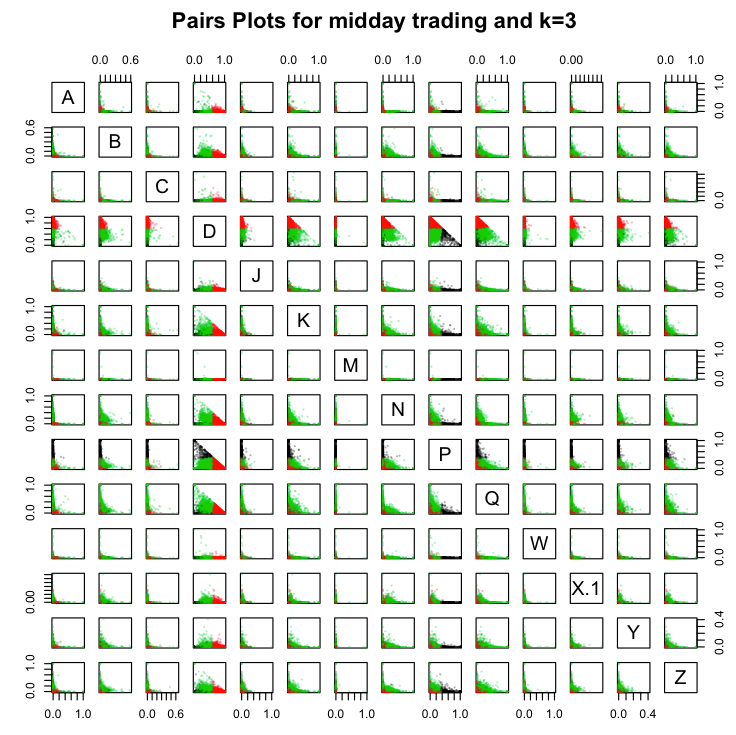
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster Name** | **Color** | **# of securities** | **Index of cluster center** | | | | | | |
| O | Black | 3676 | A | B | C | D | J | K | M |
| 0.008 | 0.028 | 0.002 | 0.408 | 0.019 | 0.082 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.086 | 0.137 | 0.143 | 0.006 | 0.003 | 0.012 | 0.068 |
| D | Red | 2895 | A | B | C | D | J | K | M |
| 0.003 | 0.009 | 0.002 | 0.780 | 0.008 | 0.037 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.030 | 0.053 | 0.049 | 0.001 | 0.001 | 0.005 | 0.023 |

*Table 17: Cluster Centers for Midday Trading and K=2*

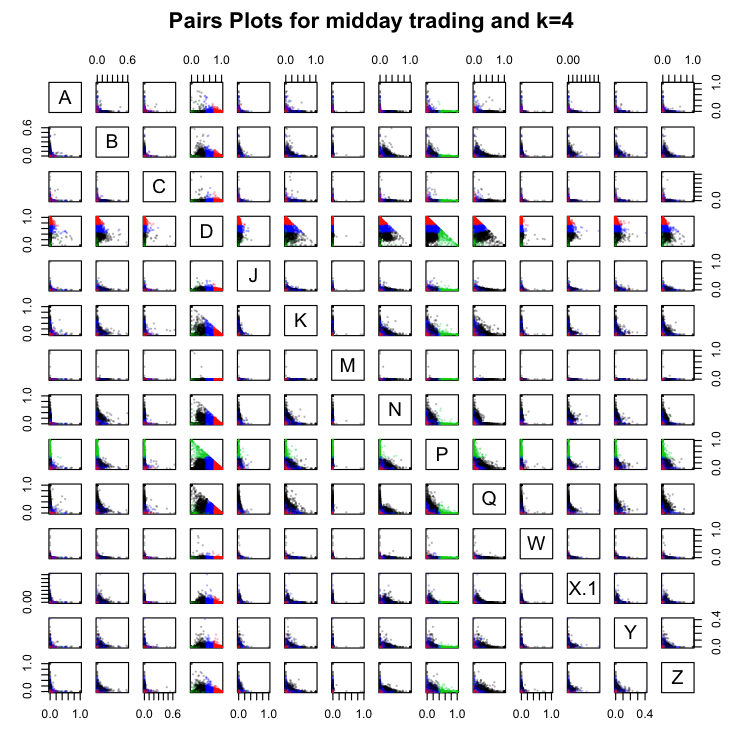
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster Name** | **Color** | **Top 5 securities based on total trading volume** | **Statistics for Total Trading Volume** | | |
| O | Black | BAC, EWJ, INTC, S, EEM | Max | Min | Average |
| 82568510 | 100 | 635881.4 |
| D | Red | GE, UVXY, NOK, ZNGA, FB | Max | Min | Average |
| 36675223 | 100 | 211925.3 |

*Table 18: Trading Volume of Clusters for Midday Trading and K=2*

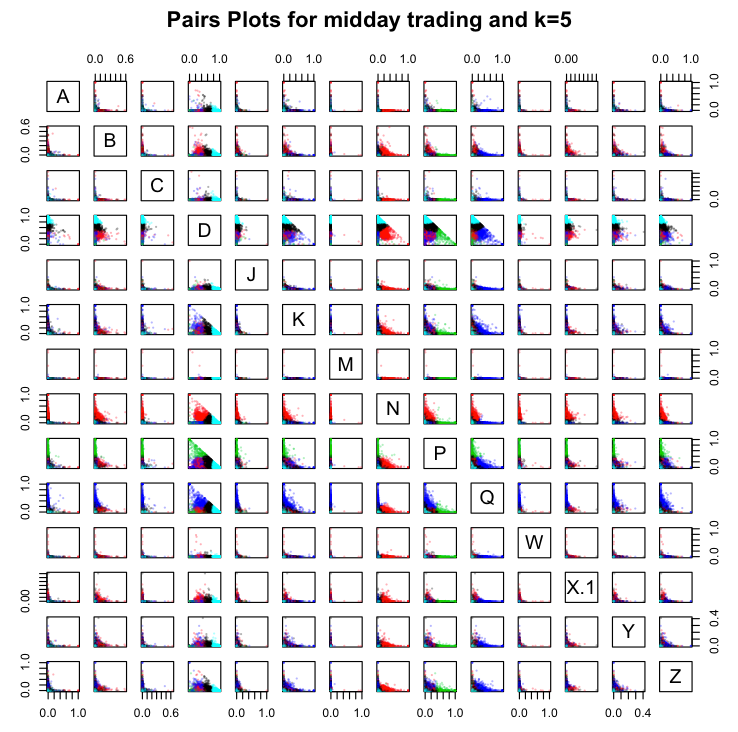




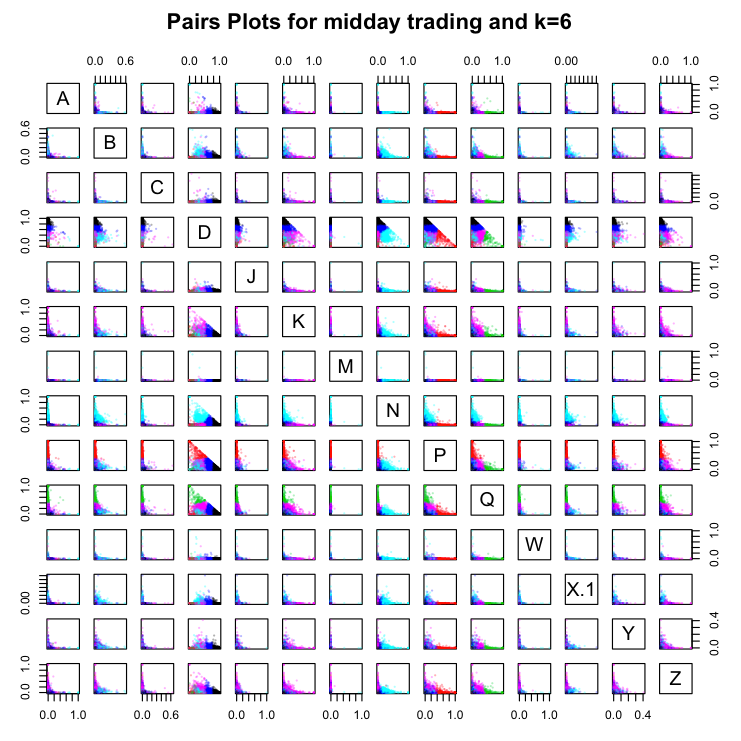
*Figure 10: Clustering plot of midday trading and k=3*



*Figure 11: Clustering plot of midday trading and k=4*



*Figure 12: Clustering plot of midday trading and k=5*



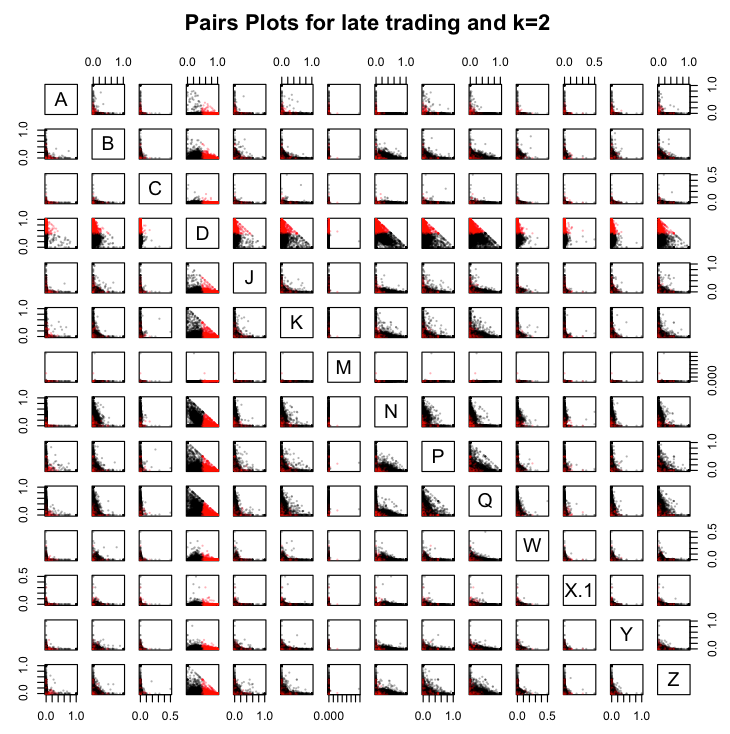
*Figure 13: Clustering plot of midday trading and k=6*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster Name** | **Color** | **# of securities** | **Statistics for Trading Volume** | | | | | | |
| D1 | Black | 1397 | A | B | C | D | J | K | M |
| 0.002 | 0.004 | 0.001 | 0.892 | 0.004 | 0.018 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.014 | 0.028 | 0.022 | 0.001 | 0.000 | 0.002 | 0.10 |
| P | Red | 235 | A | B | C | D | J | K | M |
| 0.003 | 0.003 | 0.000 | 0.242 | 0.007 | 0.023 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.008 | 0.642 | 0.040 | 0.000 | 0.000 | 0.002 | 0.029 |
| Q | Green | 154 | A | B | C | D | J | K | M |
| 0.009 | 0.006 | 0.002 | 0.232 | 0.005 | 0.043 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.000 | 0.051 | 0.609 | 0.002 | 0.000 | 0.005 | 0.034 |
| D2 | Blue | 2180 | A | B | C | D | J | K | M |
| 0.004 | 0.016 | 0.002 | 0.638 | 0.014 | 0.058 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.044 | 0.087 | 0.08 | 0.003 | 0.000 | 0.008 | 0.042 |
| N | Light Blue | 1286 | A | B | C | D | J | K | M |
| 0.006 | 0.042 | 0.001 | 0.379 | 0.020 | 0.067 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.212 | 0.099 | 0.085 | 0.010 | 0.000 | 0.013 | 0.059 |
| O | Purple | 1319 | A | B | C | D | J | K | M |
| 0.005 | 0.024 | 0.004 | 0.410 | 0.021 | 0.121 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.010 | 0.109 | 0.185 | 0.004 | 0.000 | 0.015 | 0.091 |

*Table 19: Cluster Centers for Midday Trading and K=6*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster Name** | **Color** | **Top 5 securities based on total trading volume** | **Statistics for Total Trading Volume** | | |
| D1 | Black | CLSN, SPLV, CPRX, PAL, CYTK | Max | Min | Average |
| 4908841 | 100 | 62243.3 |
| P | Red | RSX, EWM, EWS, EWY, EWP | Max | Min | Average |
| 1884745 | 100 | 53304 |
| Q | Green | SPPI, MENT, SZYM, ZOLT, CXM | Max | Min | Average |
| 945750 | 100 | 29416.56 |
| D2 | Blue | BAC, GE, INTC, S, F | Max | Min | Average |
| 82568510 | 200 | 610300.5 |
| N | Light-Blue | RAD, PBR, EMC, XRX, HPQ | Max | Min | Average |
| 17492503 | 100 | 566705.7 |
| O | Purple | EWJ, EEM, XLF, MSFT, CSCO | Max | Min | Average |
| 31646057 | 100 | 597251.4 |

*Table 20: Trading Volume of Clusters for Midday Trading and K=6*



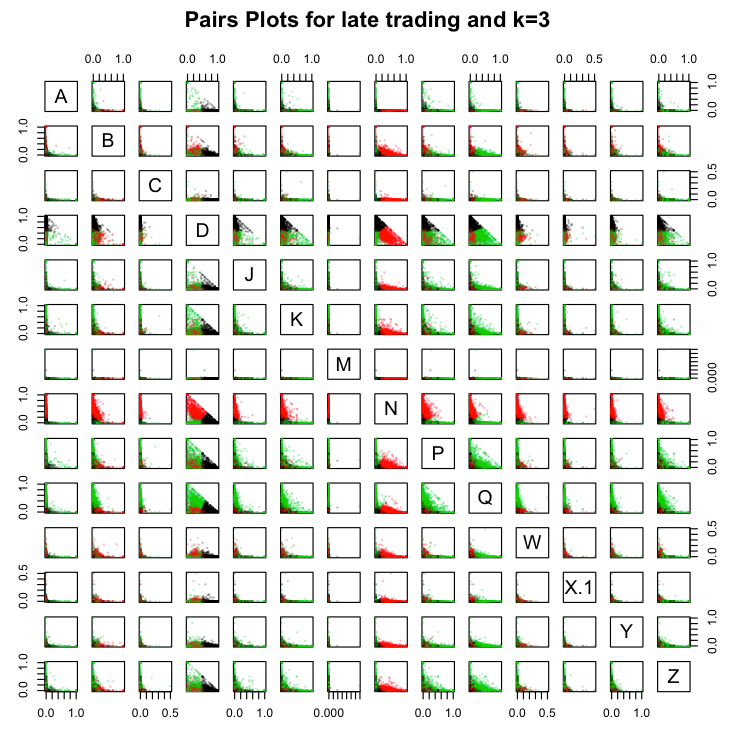
*Figure 14: Clustering plot of late trading and k=2*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster Name** | **Color** | **# of securities** | **Index of cluster center** | | | | | | |
| O | Black | 3606 | A | B | C | D | J | K | M |
| 0.009 | 0.0589 | 0.002 | 0.270 | 0.039 | 0.068 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.168 | 0.108 | 0.169 | 0.009 | 0.003 | 0.021 | 0.076 |
| D | Red | 1843 | A | B | C | D | J | K | M |
| 0.003 | 0.025 | 0.001 | 0.735 | 0.0219 | 0.032 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.033 | 0.042 | 0.057 | 0.004 | 0.001 | 0.010 | 0.034 |

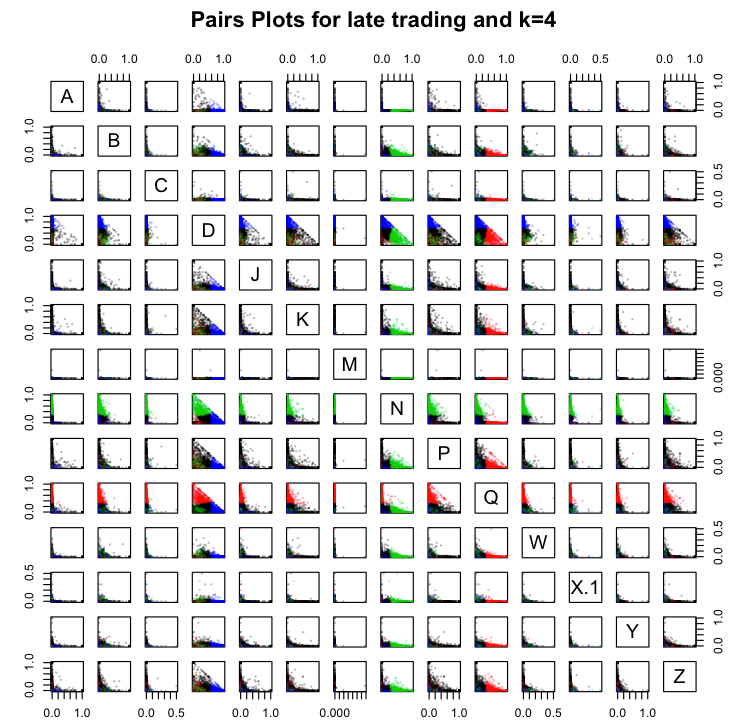
*Table 21: Cluster Centers for Late Trading and K=2*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster Name** | **Color** | **Top 5 securities based on total trading volume** | **Statistics for Total Trading Volume** | | |
| O | Black | XLF, SIRI, GE, EEM, S | Max | Min | Average |
| 4583970 | 100 | 50732.18 |
| D | Red | BAC, DELL, EWJ, NOK, RAD | Max | Min | Average |
| 5781049 | 100 | 40139.36 |

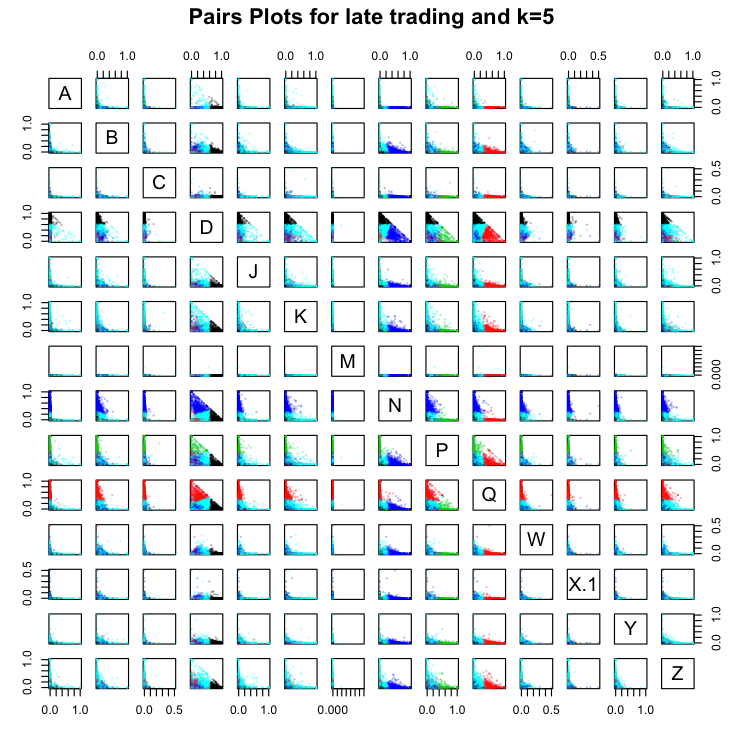
*Table 22: Trading Volume of Clusters for Late Trading and K=2*



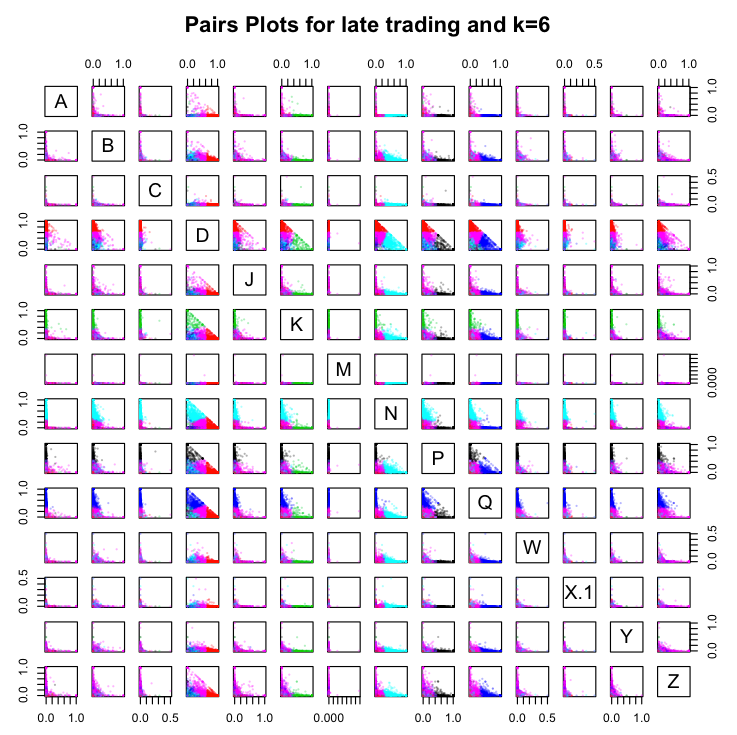
*Figure 15: Clustering plot of late trading and k=3*



*Figure 16: Clustering plot of late trading and k=4*



*Figure 17: Clustering plot of late trading and k=5*



*Figure 18: Clustering plot of late trading and k=6*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster Name** | **Color** | **# of securities** | **Statistics for Trading Volume** | | | | | | |
| P | Black | 204 | A | B | C | D | J | K | M |
| 0.004 | 0.010 | 0.001 | 0.129 | 0.019 | 0.016 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.011 | 0.717 | 0.050 | 0.000 | 0.000 | 0.007 | 0.035 |
| D1 | Red | 1155 | A | B | C | D | J | K | M |
| 0.003 | 0.014 | 0.000 | 0.845 | 0.013 | 0.021 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.025 | 0.026 | 0.029 | 0.002 | 0.001 | 0.006 | 0.017 |
| K | Green | 153 | A | B | C | D | J | K | M |
| 0.013 | 0.008 | 0.002 | 0.175 | 0.014 | 0.653 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.020 | 0.036 | 0.052 | 0.002 | 0.000 | 0.007 | 0.019 |
| Q | Blue | 644 | A | B | C | D | J | K | M |
| 0.001 | 0.032 | 0.001 | 0.220 | 0.021 | 0.041 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.004 | 0.070 | 0.531 | 0.005 | 0.001 | 0.014 | 0.060 |
| N | Light Blue | 1000 | A | B | C | D | J | K | M |
| 0.000 | 0.059 | 0.002 | 0.227 | 0.023 | 0.038 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.454 | 0.068 | 0.055 | 0.010 | 0.004 | 0.017 | 0.045 |
| D2 | Purple | 2293 | A | B | C | D | J | K | M |
| 0.014 | 0.070 | 0.002 | 0.407 | 0.054 | 0.050 | 0.000 |
| N | P | Q | W | X | Y | Z |
| 0.077 | 0.075 | 0.117 | 0.011 | 0.003 | 0.026 | 0.097 |

*Table 21: Cluster Centers for Late Trading and K=6*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster Name** | **Color** | **Top 5 securities based on total trading volume** | **Statistics for Total Trading Volume** | | |
| P | Black | XLF, EWZ, IYE, VGK, NKY | Max | Min | Average |
| 4583970 | 100 | 34585.83 |
| D1 | Red | UVXY, GTAT, SDS, FB, FITB | Max | Min | Average |
| 796093 | 100 | 14430.99 |
| K | Green | DGAZ, UCO, IVAN, VTG, MITK | Max | Min | Average |
| 90070 | 100 | 4717.19 |
| Q | Blue | NTAP, CTSH, AINV, SLM, PDLI | Max | Min | Average |
| 387205 | 100 | 12474.14 |
| N | Light-Blue | JPM, AVP, MET, PG, USB | Max | Min | Average |
| 805783 | 100 | 33429.56 |
| D2 | Purple | BAC, SIRI, GE, DELL, EEM | Max | Min | Average |
| 5781049 | 100 | 83300.98 |

*Table 22: Trading Volume of Clusters for Late Trading and K=6*