Summary Report 3

SOI Team for price prediction

Bad News:

We tested the concurrent result of using quotes mid points to calculate the bucket return. Compared to last week's best results, new model produces much weaker results. We also add a time delay to all quotes, and the results are still not satisfactory.

Given those results, and the computational capacity limit of out laptops, we decide to postpone the testing of using EMA of quotes to classify trade's BUY/SELL. The difficulty for using EMA of quotes is the need to pre-process all the quotes and trades before parsing them. If we only need the mid-points of the recent quotes, the process gets much simpler and computational time only doubles; yet for when first calculating EMA of quotes, then feeding into the parser, the pre-processing alone can consume up so many time. For details, the implementation is wrapped up as a function filter_trades_quotes_EMA in parser.csv. More importantly, based on the results we have thus far obtained, we find that using quotes only made the results worse (substantially worse that is), the results are presented below.

Use quotes mid point as bucket return proxy, and delay all the quotes by alpha millisecond, filter out large trades (>10000).
The parameters ranges are as the following:

	Start	End	Step
Bucket Size	1000	10000	1000
Time Bin	30	180	30
Delay Time(s)	0	0.1	0.005

The results are depressing:

(bucketVol_time_bin_delay)	R^2
10000 _ 150 _ 0.04	14.8%
10000 _ 150 _ 0.09	14.4%
7000 _ 150 _ 0.09	14.3%
10000 _ 150 _ 0.055	14.1%
8000 _ 150 _ 0.06	13.9%
9000 _ 150 _ 0.07	13.7%
9000 _ 150 _ 0.055	13.3%
10000 _ 180 _ 0.055	13.1%
10000 _ 150 _ 0.045	13.0%
8000 _ 180 _ 0.06	13.0%
8000 _ 150 _ 0.09	12.6%
10000 _ 150 _ 0.035	12.4%
10000 _ 180 _ 0.04	11.9%
10000 _ 150 _ 0.1	11.5%
7000 _ 150 _ 0.025	11.4%
8000 _ 150 _ 0.045	11.3%

9000 _ 150 _ 0.075	11.3%
7000 _ 180 _ 0.09	11.2%

The case when delay time equals 0 can be used to compare with our previous implementation (when we ignore quotes completely and only implemented a Lee/Ready type of classification based on transaction prices). And if you remember from last week's report, some concurrent regression yielded R^2 as large as 41.3%; but if we substitute quotes mid-points and use them to calculate bucket returns, the best concurrent regression only yielded a 14% R^2.

Good News:

So far, we conclude that the Lee-Ready concurrent model involving three parameters: bucket size, time bin, and threshold is the best model. And after taking a closer look at the data, we got inspired with a new direction to improve the Lee-Ready concurrent model. We found that the latency variable is the difference between time (the variable in the data, i.e. the time when the corresponding row of data received) and exchange time.

lantency*0.001 = time – exchange time

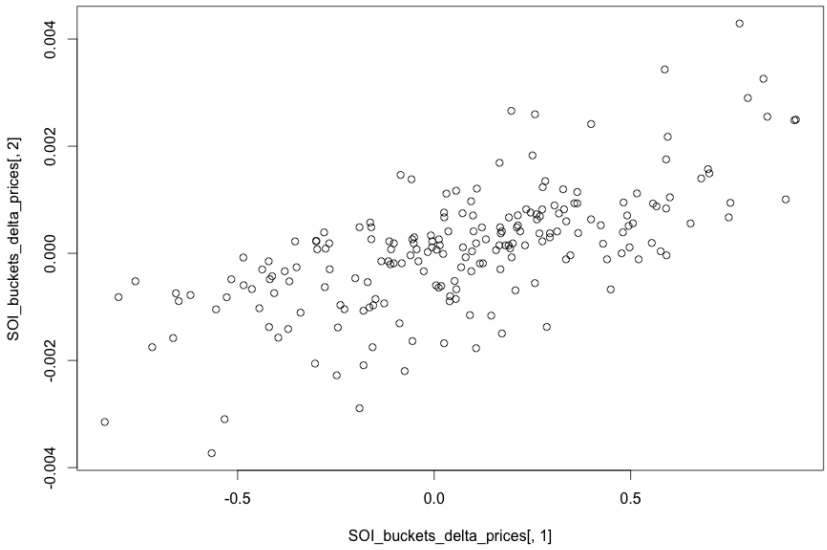
➔ Thus, exchange time = time – latency*0.001 (1)

Use formula (1) to adjust the time variable, we can get the actual time that each quote and trade got posted in the Exchanges. We rerun the optimization of the Lee Ready model (Bucket Size, Time bin, threshold) and use the trades return as bucket return proxy. (so the only difference between this set of runs and last week's run was adjusting the trade time for latency). And we have obtained better R^2.

For best Lee-Ready results, slightly improved the R^2 from 41.3% to 52.18%.

(Bucket,Time bin, Trade EX)	Adj-R2
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%
13000 _ 135 _ 1000	47.25%
11000 _ 135 _ 1000	47.18%
10000 _ 135 _ 1000	47.11%
13000 _ 135 _ 2000	47.10%
15000 _ 60 _ 1000	47.04%
14000 _ 150 _ 1000	46.40%
10000 _ 150 _ 1000	46.40%
15000 _ 120 _ 1000	46.14%
15000 _ 45 _ 1000	46.06%
2000 _ 30 _ 2000	46.04%
8000 _ 135 _ 1000	44.87%
11000 _ 150 _ 1000	44.72%
13000 _ 150 _ 1000	44.09%

13000 _ 45 _ 1000	42.93%



but the lagged prediction result is still not very satisfactory

Contingency analysis

	Early	Midday	Late
D1	3	1265	24
D2	7828	47400	1715
O	5149	28287	840
Q	1423	9523	494

Table 1 (Contingency Analysis for # of Transaction)

Cramer's V is 0.04832802 – weak relationship

Instead of making a frequency table, I attach a mosaic plot instead.

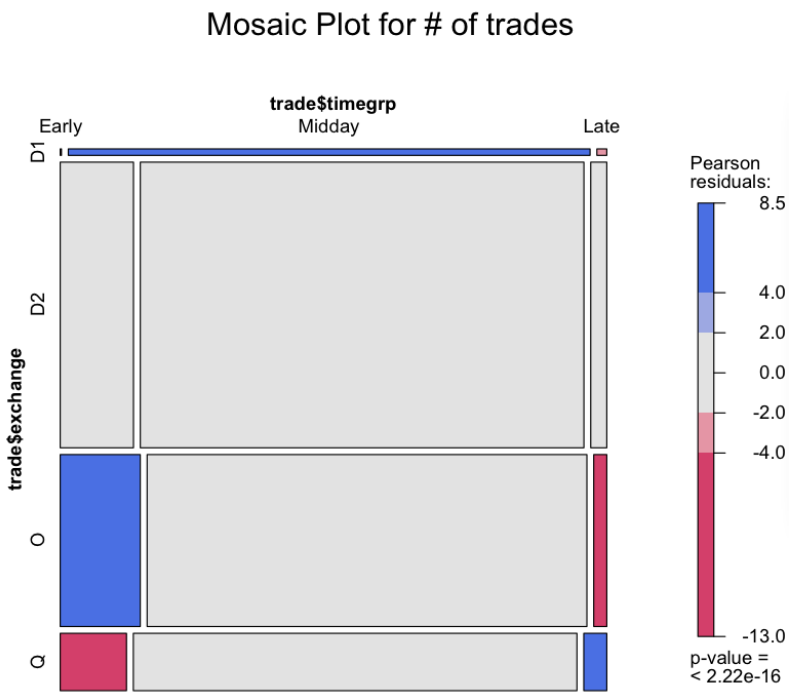


Figure 1

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

In this mosaic plot, most of the cells are not significant.

	Early	Midday	Late
D1	771	289821	5135
D2	1482682	11042741	368502
O	633770	3818988	98148
Q	180865	1411640	83055

Table 2 (Contingency Analysis for # of shares)

Cramer's v = 0.04946721

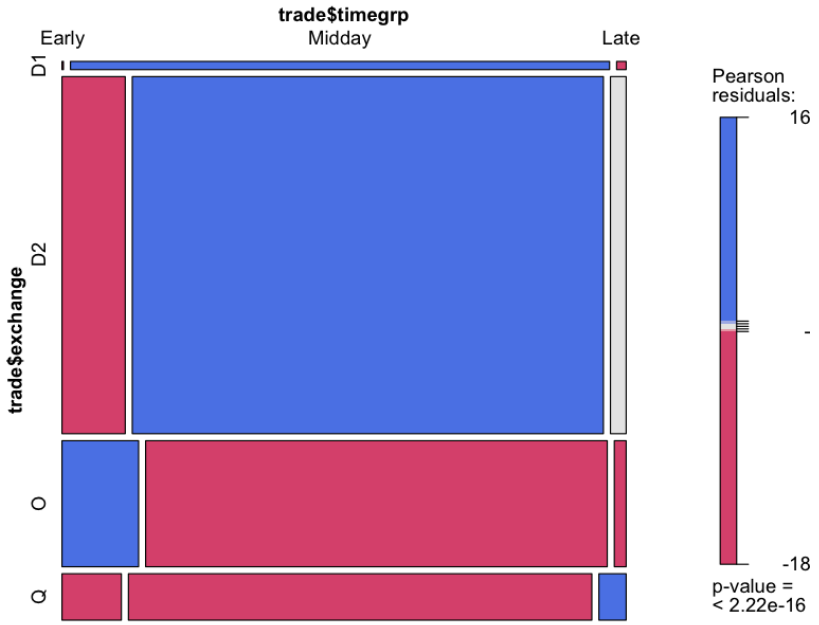


Figure 2

If we focus on the number of shares, the mosaic plot will be completely different. Most of the shares are traded in midday in D as "block trades". In general, Apple stock traded in midday is more likely to be traded in exchange D; AAPL traded in early morning is more likely to be trades in other exchange; AAPL traded in late time is more likely to be traded in the primary exchange.

However, as we use Pearson residual (chi-square) test here, number of share MAY magnify chi-square test. So we conclude that the result is not significant, and the result of mosaic plot is not significant as well.

If we compare D2 vs non-D2, the result is not significant at all.

	Early	Midday	Late
Non-D2	6575	39075	1358
D2	7828	47400	1715

Table 3 (Contingency Analysis for # of Transaction)

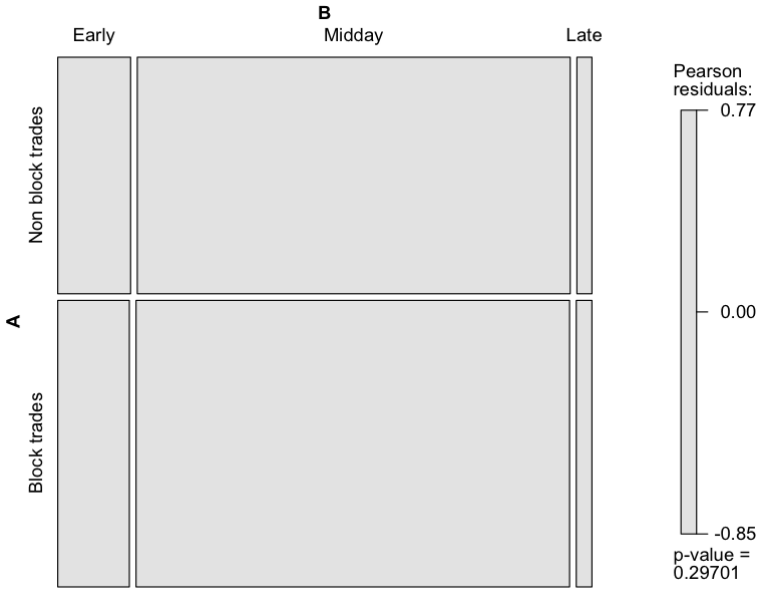


Figure 3

	Early	Midday	Late
Non-D2	815406	5520449	186338
D2	1423	9523	494

Table 2 (Contingency Analysis for # of shares)

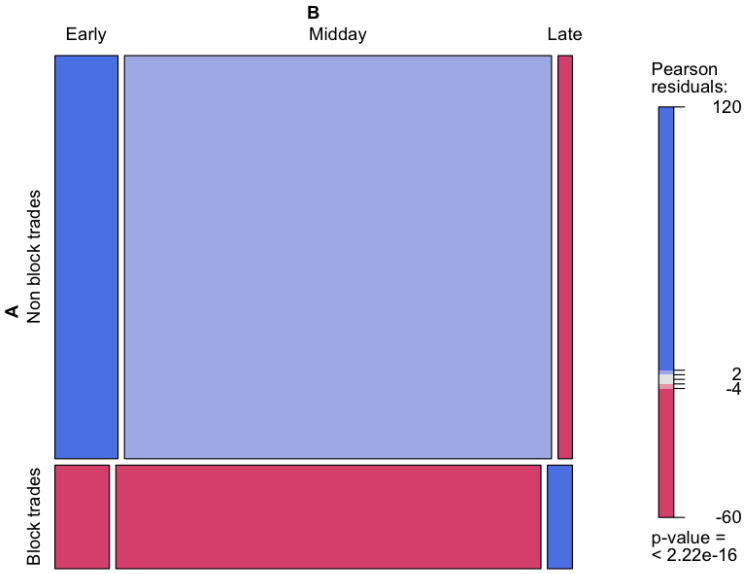


Figure 4

Cramer's v for table 3 and table 4 are 0.004832872 and 0.0507463. So for table 4, it we can say there are some weak relationship between time and location
From the mosaic plot, we can see that non-D2 is more likely to be traded in the morning, and D2 is more likely to be traded in the late 10 minutes. Again, result is not significant.

Multinomial Logistic Regression

$$\textbf{Model: } \ln\frac{P(exchange=?)}{P(excahnge=B)} = \beta_0 + \beta_1 \times I_{Midday} + \beta_2 \times I_{Late}$$

	Intercept	I_{Midday}	I_{Late}
D2	7.866580	-4.243032	-3.597384
O	7.447675	-4.340347	-3.892246
Q	6.161640	-4.143005	-3.137077

Table 5 (Coefficients of MLR model)

The negative log-likelihood for this model is 102959.091624, which is much better than the previous ones (more than 150000)

Interpretation:

An example of interpretation of coefficient of indicator variable:

$$\ln\frac{P(exchange=D2)}{P(excahnge=D1)} = 7.866580 - 4.243032 \times I_{Midday} - 3.597384 \times I_{Late}$$

- The log odds of trade happening in exchange D as D2 vs D1 will decrease by 4.243032(b_12) if moving from time=early to time=midday
- The log odds of trade happening in exchange D as D2 vs D1 will decrease by 3.597384(b_13) if moving from time=early to time=late
- The result displays trades happening later in a day in exchange D is more likely to be classified as D1.

Clustering

Please see "Clustering plots" folder

Three data files represent the trading volume for each stock through all stock exchanges in the first ten minutes, in the last ten minutes, and in the remaining time. Each column represents one stock exchange, each row represents one stock, and the number in cell represents the trading volume of the stock in the corresponding stock exchange. We standardize the data by dividing each element by row total.

What we do here is trying to cluster all the stock in order to minimize the within-group variation and maximize between-group variation. Stocks 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.

$$\mathbf{CH(K)} = \frac{\mathbf{B(K)/(K-1)}}{\mathbf{W(K)/(n-k)}}$$

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.

Interpretation of clustering plot:

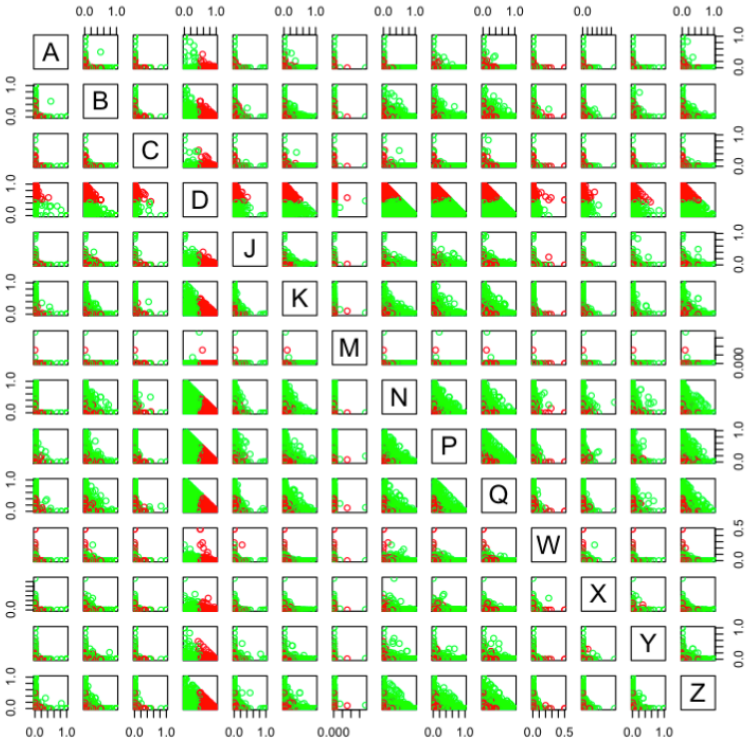


Figure 5: Clustering plot of early trading and k=2

The determent factor in this clustering is exchange D. If most of the trades of a certain stock happen in exchange D, it might be labeled red; if most of the trades are done in other exchanges, the stock will be labeled green in the plot. So the next step is to explore the features in each group.

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

Table 6 (significant factors according to clustering plots)

Compare the trading volume between two clusters.

We are comparing the mean trading volume of stocks in cluster 1 (more trades in D) and in cluster 2 (less trades in D). When check the boxplots of trading volume in each cluster, there are several outliers, which may inflate the mean. So when we delete these outliers when computing the average trading volume.
For the first 10 minutes, BAC is an outlier regarding trading volume (1941695), so we delete BAC, and check the boxplots of cluster 1 and cluster 2. The distribution of the two clusters look similar.
Similarly, we delete BAC and GE for midday; BAC and XLF are deleted for the last ten minutes.

Trading Volume	Overall mean	Mean of cluster 1	Mean of cluster 2
Early	30266.72	35961.27	24292.12
Midday	441649.5	201074.6	613781.8
Late	46126.72	37022.69	49474.7

Table 7 (average trading volume of each cluster)

Remember cluster 1 contains all the stocks, most of whose trades take place in exchange D. Comparing the average trading volumes, we found that in the morning, stocks with large trading volume are more likely to be traded in exchange D; but in the midday or late, it is just the opposite. This test is not robust without checking more data. We can use some other data to check whether this conclusion stands or not.

Compare Stock's Primary Exchange through All Clusters

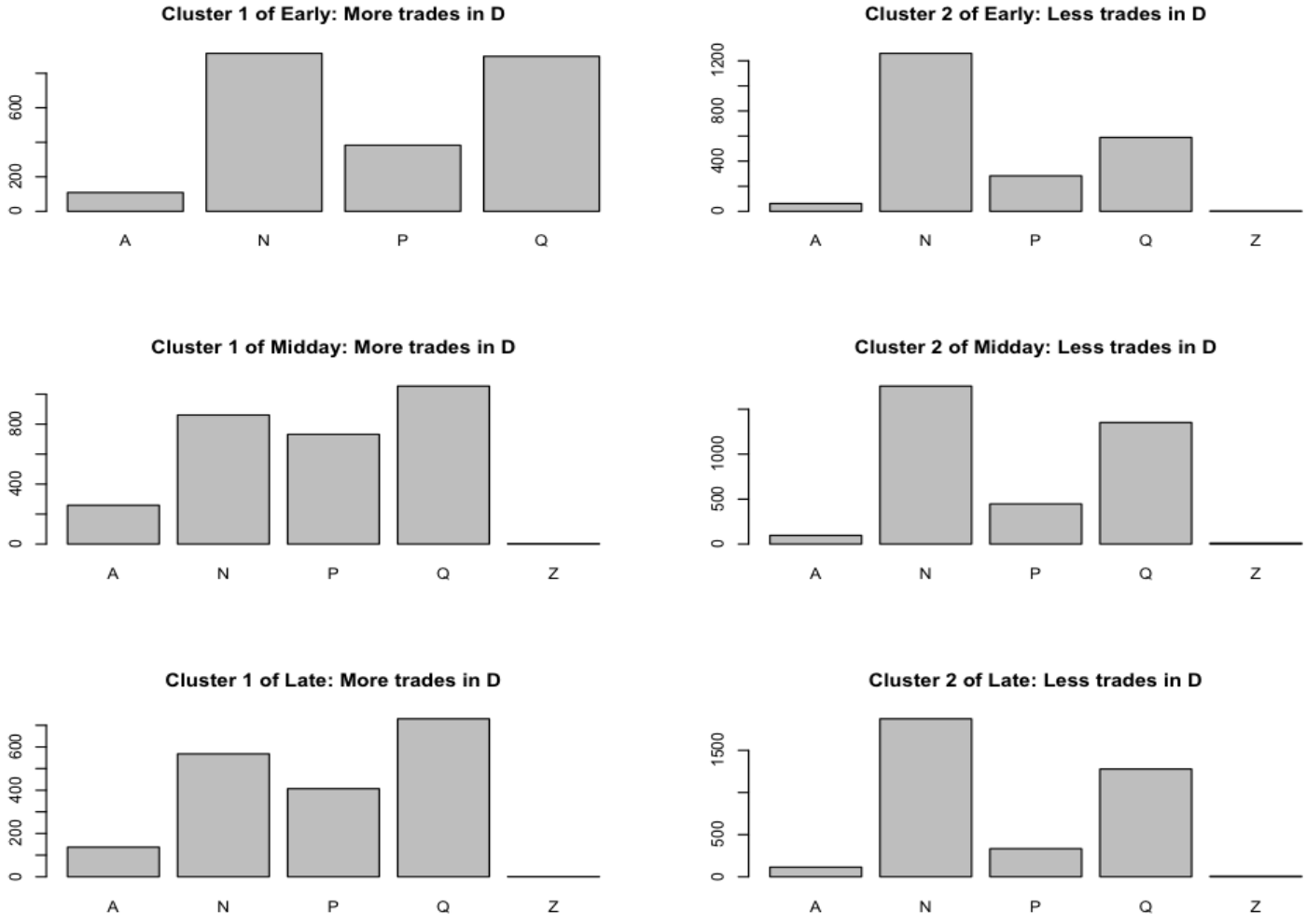


Figure 6: Barplots of Stocks' Primary Exchange in different clusters

From the barplots, the main observation is that stocks whose primary exchange is P and Q are more likely to be traded in D, especially for exchange P.

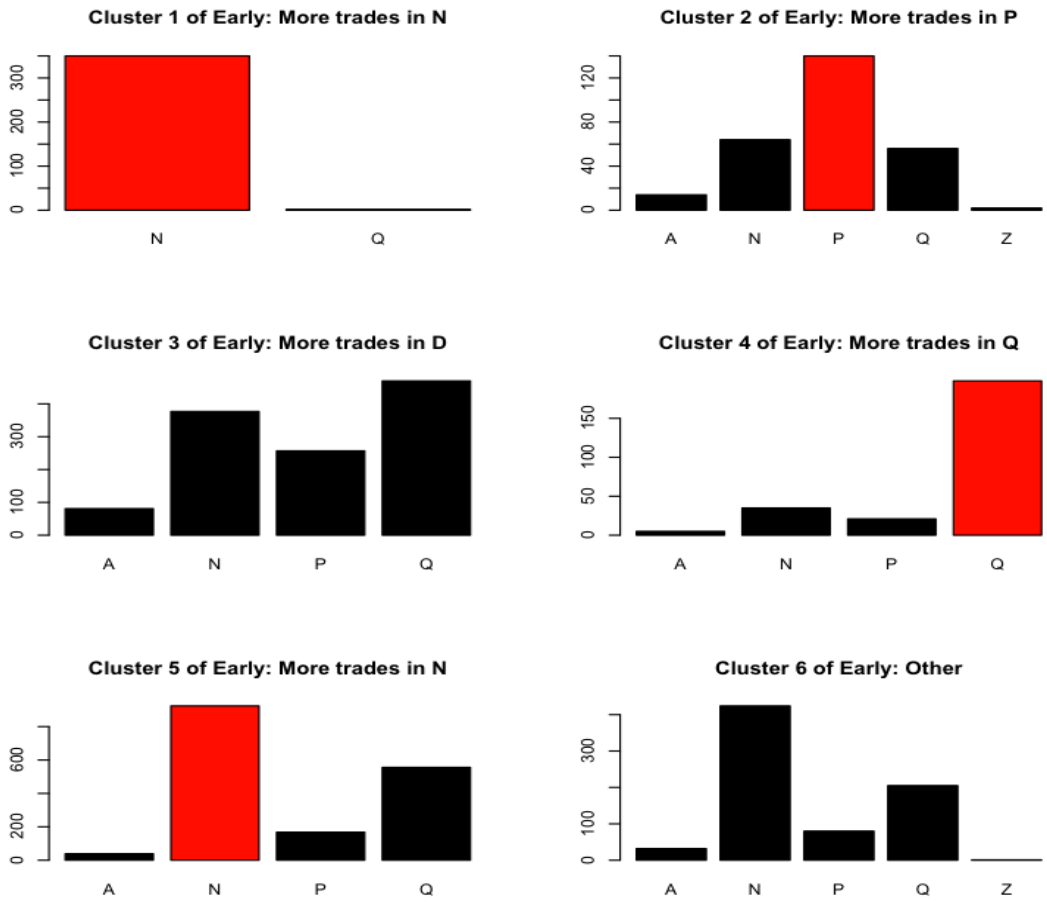


Figure 7: Barplots of Stocks' Primary Exchange in 6 different clusters in early

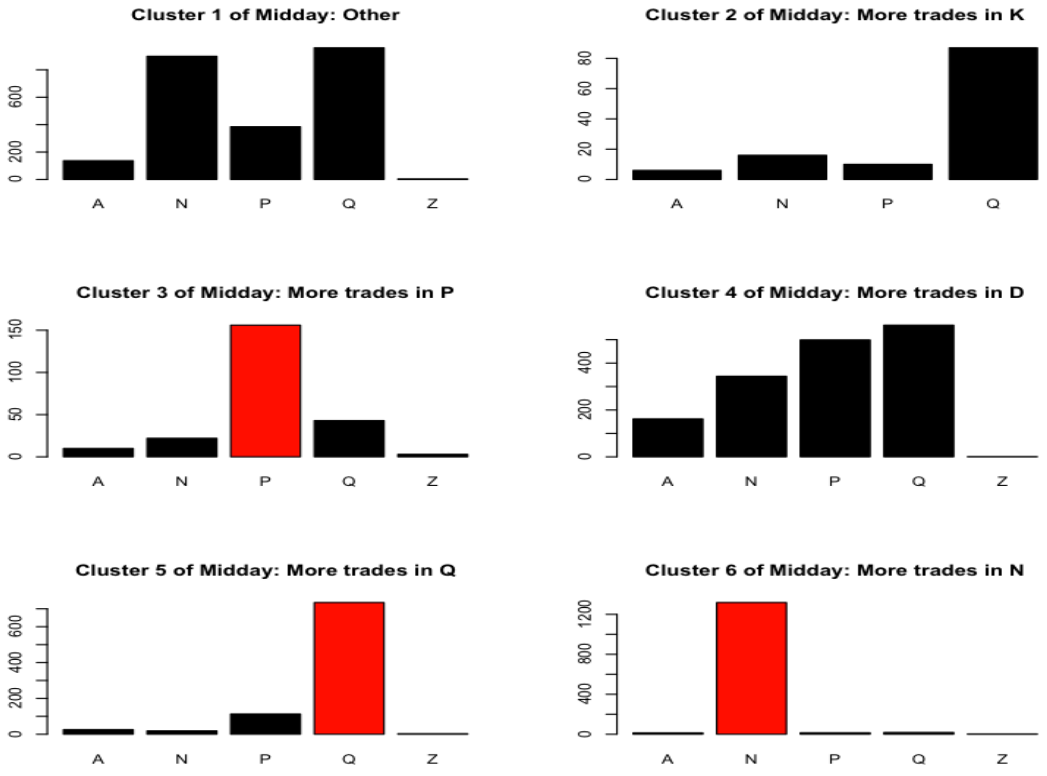


Figure 8: Barplots of Stocks' Primary Exchange in 6 different clusters in midday

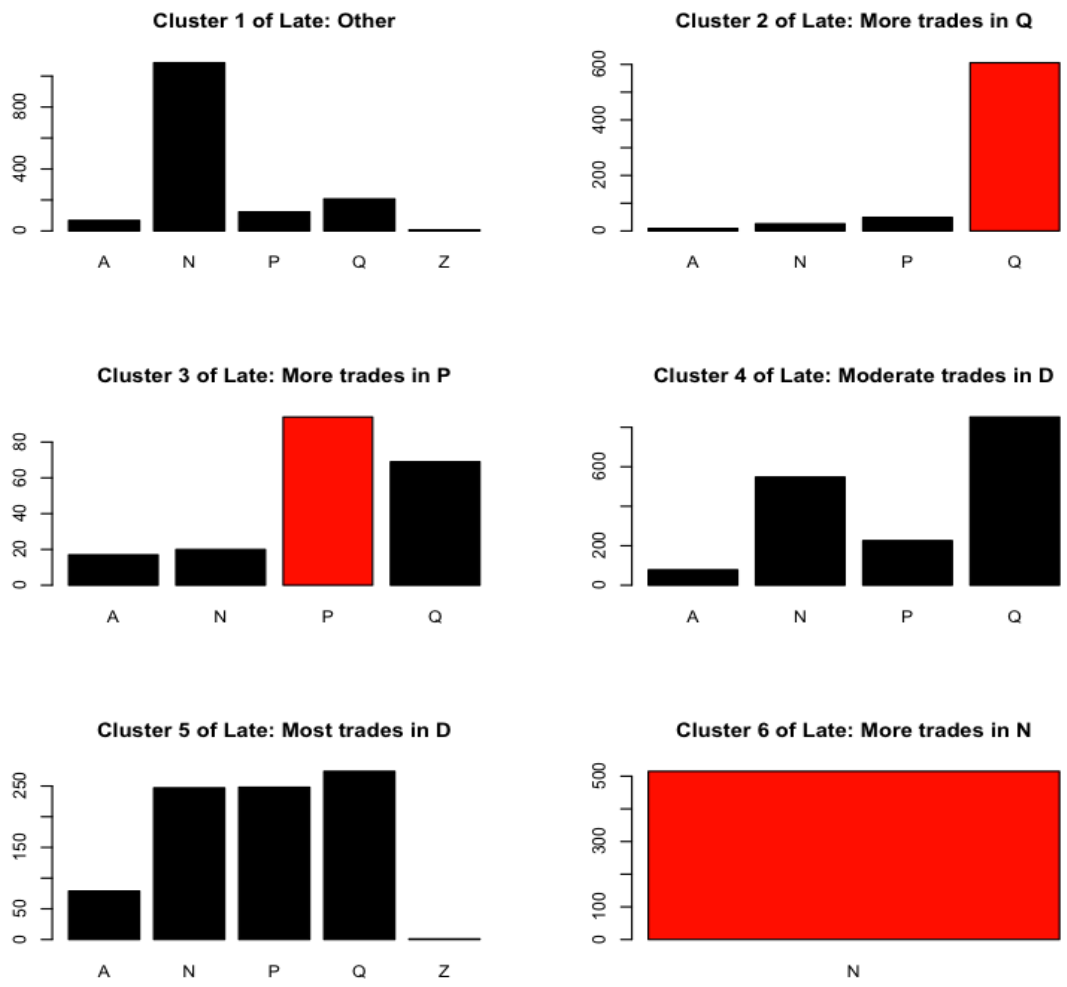


Figure 9: Barplots of Stocks' Primary Exchange in 6 different clusters in late

Observations:

- Stocks are mostly traded in their primary exchanges. This is true for all the time.
- Almost all the stocks traded at N are listed on N.
- Most of the stocks trades at Q are listed on Q.
- Q-listed stock are popular in D.

Principal Component Analysis

- Early period

	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 8 (deviation PC can explain)

First 8 principal component directions can explain about two third of the variation in the data set.

	PC1	PC2	PC3	PC4	PC5	PC6
A	0.0300749049	-0.022632141	0.067586707	-0.13995449	-0.02734999	0.16160310
B	0.1812903627	0.191238855	-0.307321061	0.14882516	-0.10480199	0.11756987
C	0.0001773848	0.011215181	-0.033677047	-0.13340959	-0.24624261	0.07997137
D	-0.7705468099	-0.009416201	-0.050649741	0.10609387	0.03544900	-0.02372358
J	0.1315173595	-0.047959778	-0.148051851	0.42271610	-0.24550113	-0.39982413
K	0.1460089757	-0.185488582	-0.236084887	-0.19139674	-0.63513917	0.20154303
M	0.0058129296	-0.049443812	-0.008742248	0.09609907	0.01671406	-0.05119527
N	0.2580023678	0.680427730	0.061098016	-0.44739540	0.04704284	-0.13144611
P	0.2991826916	-0.131013327	0.749623785	0.33793993	-0.15001424	0.13293079
Q	0.2900163673	-0.496652933	-0.352133523	-0.14259010	0.24747006	0.25876889
W	0.0716600766	0.271829169	-0.194212177	0.38749603	0.20511441	0.36614021
X	0.0967593316	0.270383643	-0.173125609	0.40508604	0.12801324	0.32004414
Y	0.1177042067	0.071699450	-0.246564571	0.23392483	-0.22349927	-0.51169930
Z	0.2555403035	-0.211188187	-0.037452272	-0.03140796	0.51773986	-0.38700202

Table 9 (First 6 PC direction)

- Midday period

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.5036	1.1842	1.09720	1.04174	1.00663	1.00180	0.99568	0.98627
Proportion of Variance	0.1615	0.1002	0.08599	0.07752	0.07238	0.07169	0.07081	0.06948
Cumulative Proportion	0.1615	0.2616	0.34764	0.42516	0.49753	0.56922	0.64003	0.70951

Table 10 (deviation PC can explain)

First 8 principal component directions can explain about 70% of the variation in the data set.

	PC1	PC2	PC3	PC4	PC5	PC6
A	0.030233780	-0.09572922	0.02348490	-0.293759364	0.839052167	-0.041708008
B	-0.351993232	0.30548640	-0.08411861	-0.032605883	0.066386685	0.007378696
C	0.004404556	-0.10330864	-0.01548344	-0.441876088	0.235209249	0.183432803
D	0.584404696	0.26987118	-0.25983253	0.152484463	-0.005577389	0.023499323
J	-0.279751165	0.09823067	-0.20526066	0.306916059	0.161669966	0.039180285
K	-0.171909188	-0.23238820	-0.16955210	-0.579851216	-0.382817234	0.082077048
M	-0.001755804	-0.01838336	0.06303498	-0.060501846	0.022097643	-0.952428364
N	-0.231252698	0.49264130	0.30960272	-0.277126304	-0.111884705	0.004144213
P	-0.135063594	-0.30312684	0.74041964	0.259471328	0.026991421	0.112714116
Q	-0.254034024	-0.47222495	-0.32279409	0.040330058	-0.044614298	-0.081464216
W	-0.100529365	0.08914919	0.06836164	-0.118853895	-0.167741363	-0.164748712
X.1	-0.268104416	0.38931499	0.01173249	-0.009921591	0.086492370	0.012429164
Y	-0.299022400	0.12629765	-0.26650839	0.123812958	0.095762769	0.039590046
Z	-0.343755830	-0.13029503	-0.16018044	0.291111188	0.055040807	-0.019814155

Table 11 (First 6 PC direction)

- Late period

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.2733	1.18242	1.10796	1.04650	1.02329	1.00706	1.00599	0.99817
Proportion of Variance	0.1158	0.09986	0.08768	0.07823	0.07479	0.07244	0.07229	0.07117
Cumulative Proportion	0.1158	0.21568	0.30336	0.38159	0.45638	0.52882	0.60111	0.67228

Table 12 (deviation PC can explain)

First 8 principal component directions can explain about two third of the variation in the data set.

	PC1	PC2	PC3	PC4	PC5	PC6
A	0.025267842	-0.07388095	0.14868443	-0.03906307	-0.506992317	0.1315433688
B	-0.370190160	0.32344861	-0.24950848	-0.01389296	0.033205596	0.1124657259
C	-0.082284221	-0.06380221	0.08727229	0.28325615	0.004974356	-0.4526286644
D	0.701201929	0.28999724	-0.21024867	-0.03968961	0.028488082	-0.0418154228
J	-0.165965006	0.03151161	-0.33439061	-0.07817134	-0.506054541	-0.1454731563
K	-0.047869336	-0.18627725	0.18520787	0.75757038	-0.263373508	-0.0622091602
M	-0.005020309	-0.05601606	0.01229862	-0.15483389	0.058023460	-0.2391206492
N	-0.309267576	0.43791535	0.51286479	-0.01281998	0.194451467	0.0007307862
P	-0.156093965	-0.32837128	0.37358242	-0.50378584	-0.227456413	0.0839901309
Q	-0.124829259	-0.56661347	-0.25434235	0.11003302	0.409642512	0.3982457466
W	-0.274534660	0.25744532	-0.18942595	0.13220687	0.001485085	0.2078555067
X.1	-0.099577064	0.13647742	0.07544086	0.04125499	0.308455461	-0.0920261461
Y	-0.272798560	0.12332801	-0.39915636	-0.03396710	-0.179020084	0.0740153347
Z	-0.196740578	-0.20533178	-0.22217951	-0.16145711	0.166272039	-0.6770597593

Table 13 (First 6 PC direction)