**Topic 2: Statistical Analysis Summary**

In topic 2, we would like to explore the factors that influence where stocks trade and develop statistical model to provide prediction on exchange location of further transaction. This summary will focus on the AAPL’s transactions in one single day (201304240), and explore the relationship between exchange location, transaction time, latency and exchange size.

There are more than 100 thousand transactions in this data set.

**Contingency table analysis:**

First, we will consider the number of transactions only, regardless of any transaction size.

|  |  |  |  |
| --- | --- | --- | --- |
| **exchange** | **early** | **midday** | **late** |
| **B** | 627 | 3697 | 113 |
| **C** | 13 | 52 | 0 |
| **D** | 7831 | 48665 | 1739 |
| **J** | 747 | 2963 | 94 |
| **K** | 761 | 5337 | 153 |
| **P** | 1785 | 9401 | 233 |
| **Q** | 1423 | 9523 | 494 |
| **W** | 5 | 8 | 0 |
| **X** | 1 | 9 | 0 |
| **Y** | 286 | 1111 | 85 |
| **Z** | 924 | 5709 | 162 |

*Table 1 (Contingency Analysis for # of Transaction)*

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

where Oi is observation, and Ei=

Cramér's V is a 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. The higher Cramér's V is, the stronger relationship between the two categorical variables.

For table 1, Cramér's V is 0.0433153, so we cannot reject the independent null hypothesis.

If we take trading size into consideration, the contingency table will be weighted by transaction size.

|  |  |  |  |
| --- | --- | --- | --- |
| **exchange** | **early** | **midday** | **late** |
| **B** | 67720 | 385534 | 11761 |
| **C** | 1400 | 5495 | 0 |
| **D** | 1483453 | 11332562 | 373637 |
| **J** | 81439 | 331072 | 10957 |
| **K** | 104999 | 927250 | 22276 |
| **P** | 236777 | 1322190 | 27904 |
| **Q** | 180865 | 1411640 | 83055 |
| **W** | 500 | 1000 | 0 |
| **X** | 100 | 1400 | 0 |
| **Y** | 29482 | 119324 | 8587 |
| **Z** | 111353 | 16663 | 725723 |

*Table 2 (Contingency Analysis for Transaction Size)*

For table 2, Cramér's V is 0.6954657, so we can reject the independent null hypothesis and conclude that different time sessions in one single day have great influence on where AAPL trade regarding the trading size.

**Multinomial Logistic Regression:**

Multinomial logistic regression is used to model nominal outcome variables, in which the log odds of the outcomes are modeled as a linear combination of the predictor variables.

First we will only include latency as predictor variable.

**Model 1: In = +**

|  |  |  |
| --- | --- | --- |
|  | Intercept | latency |
| C | -4.2524111 | 0.0032864160 |
| D | 2.5821547 | -0.0012179358 |
| J | -0.1374674 | -0.0028861879 |
| K | 0.3429093 | -0.0000144665 |
| P | 0.9502796 | -0.0007757768 |
| Q | 0.9451624 | 0.0002908742 |
| W | -5.7622252 | -0.0175182751 |
| X | -5.8802861 | -0.0703286371 |
| Y | -1.0943394 | -0.0003443443 |
| Z | 0.4206977 | 0.0007786724 |

*Table 3 (Coefficients for MLR Model 1)*

The interpretation of coefficients in MLR is like that in logistic regression. We interpret log odd ratio based on the probability trade happens in exchange B. For example,

**In = -4.2524111+**

* A one-unit increase in variable latency is associated with the increase in the log odds of trade happening in exchange C vs in exchange B in the amount of 0.003286416(b\_1)

All the coefficients in *Table 3* are unstandardized, so we display standardized coefficients (z-score) in *Table 4* and p-value in *Table 5* for easier comparison.

|  |  |  |
| --- | --- | --- |
|  | Intercept | latency |
| C | -33.258293 | 1.47960969 |
| D | 161.793259 | -2.31365392 |
| J | -6.053086 | -3.02468662 |
| K | 17.049271 | -0.02214783 |
| P | 52.411480 | -1.27119011 |
| Q | 52.164530 | 0.50112758 |
| W | -18.332049 | -0.36874138 |
| X | -10.577370 | -0.40115794 |
| Y | -35.587587 | -0.33122658 |
| Z | 21.279774 | 1.27225078 |

*Table 4 (Standardized Coefficients for MLR Model 1)*

|  |  |  |
| --- | --- | --- |
|  | Intercept | latency |
| C | 0.000000e+00 | 0.0347443600 |
| D | 0.000000e+00 | 0.0051716783 |
| J | 3.552423e-10 | 0.0006222272 |
| K | 0.000000e+00 | 0.2455825073 |
| P | 0.000000e+00 | 0.0509152562 |
| Q | 0.000000e+00 | 0.1540703349 |
| W | 0.000000e+00 | 0.1780801248 |
| X | 0.000000e+00 | 0.1720759617 |
| Y | 0.000000e+00 | 0.1851183369 |
| Z | 0.000000e+00 | 0.0508210076 |

*Table 5 (P-value of Coefficients for MLR Model 1)*

If we use 0.05 as a threshold, all the intercept terms are significant in model 1. For the coefficient of latency, many terms are not statistically significant, so we can interpret that one-unit change of latency may not change the corresponding log odds ratio much.

Then we try to include trading time in the model as well.

**Model 2: In = +++**

Because time is a three-level categorical variable, we use and as indicator variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Intercept | latency | IMidday | ILate |
| C | -3.6955989 | 4.163702e-03 | -0.66507940 | -12.93180972 |
| D | 2.5323278 | -1.397311e-03 | 0.05377577 | 0.22116870 |
| J | 0.1877503 | -2.770358e-03 | -0.39233547 | -0.32547263 |
| K | 0.1940166 | -2.998229e-05 | 0.17287337 | 0.10228680 |
| P | 1.0506926 | -6.628618e-04 | -0.11304212 | -0.32291597 |
| Q | 0.8208991 | 1.075283e-05 | 0.12542922 | 0.64693217 |
| W | -4.4588692 | -3.107709e-02 | -1.53875388 | -2.71022721 |
| X | -6.0624911 | -1.786172e-01 | 0.61071315 | 1.10797247 |
| Y | -0.7787499 | -7.450235e-04 | -0.41433677 | 0.50486300 |
| Z | 0.3844264 | 8.202302e-04 | 0.04470414 | -0.05727289 |

*Table 6 (Coefficients for MLR Model 2)*

An example of interpretation of coefficient of indicator variable:

**In = -3.6955989 +4.163702e-03-0.66507940IMidday -12.93180972 ILate**

* The log odds of trade happening in exchange C vs exchange B will decrease by 0.66507940(b\_12) if moving from time=early to time=midday
* The log odds of trade happening in exchange C vs exchange B will decrease by 12.93180972(b\_13) if moving from time=early to time=late

Table 7 displays the p-value of all the coefficients, which helps us to determine significance.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Intercept | latency | IMidday | ILate |
| C | 0.000000e+00 | 0.0144295031 | 5.634572e-03 | 0.000000e+00 |
| D | 0.000000e+00 | 0.0022197136 | 5.773312e-02 | 9.019649e-03 |
| J | 1.369113e-04 | 0.0009270451 | 1.099193e-11 | 7.405066e-03 |
| K | 8.276320e-05 | 0.2409712771 | 7.259164e-04 | 1.124838e-01 |
| P | 0.000000e+00 | 0.0705172228 | 6.179093e-03 | 2.266043e-03 |
| Q | 0.000000e+00 | 0.2463730307 | 3.824266e-03 | 4.358139e-09 |
| W | 0.000000e+00 | 0.1784211057 | 8.135745e-04 | 0.000000e+00 |
| X | 3.522183e-13 | 0.1279893605 | 9.549316e-02 | 2.292179e-02 |
| Y | 0.000000e+00 | 0.1213065454 | 4.055022e-08 | 4.186038e-04 |
| Z | 3.042011e-14 | 0.0467344673 | 1.059787e-01 | 1.668944e-01 |

*Table 7 (P-value of Coefficients for MLR Model 2)*

**Conclusion:**

The negative log-likelihood for model 2 (153674.874324) is lower than that of model 1 (153858.548522), which means model 2 is more desirable. From Table 7, we can see that almost all the coefficients for time indicator variables are statistically significant, which means when the stock trades have significant influence on where the stock trades.

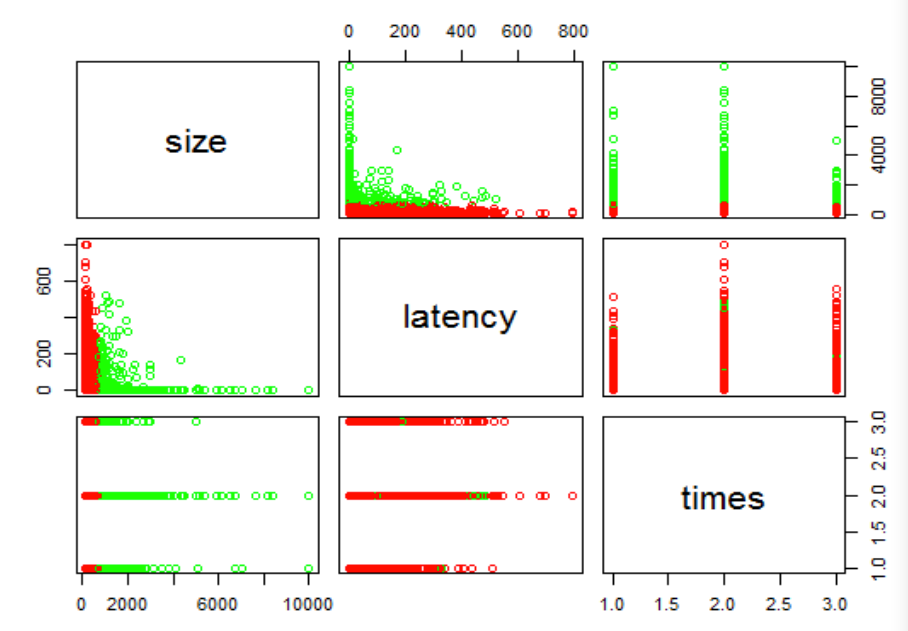
**Clustering Analysis:**

K-means clustering is one of the most popular clustering methods, however, it requires all the variables to be numeric. So I transform time variables from categorical (early, midday, late) to ordinal (1, 2, 3), then cluster based on time, size and latency.

CH value is a good criterion to determine the goodness of a clustering.

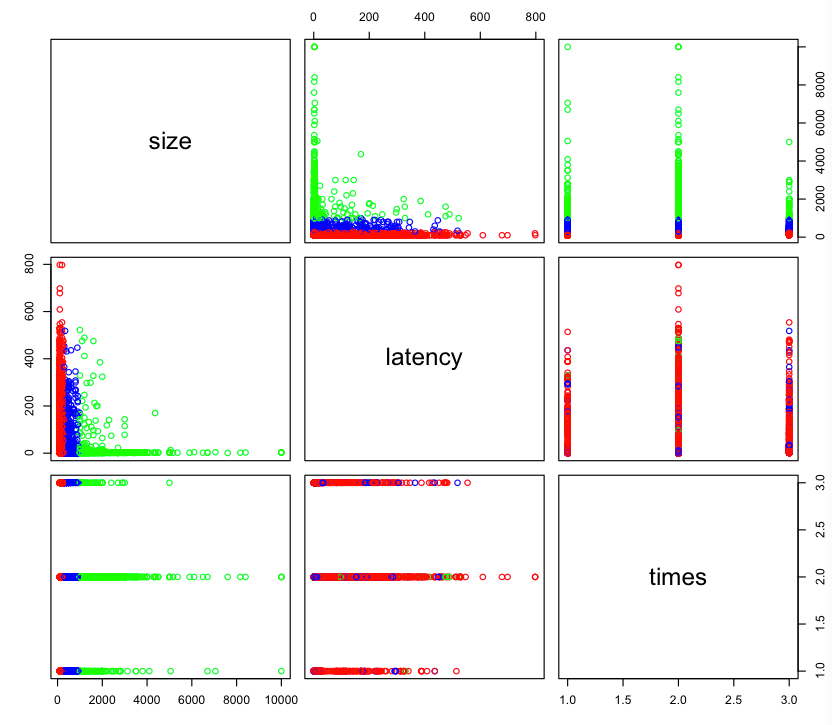
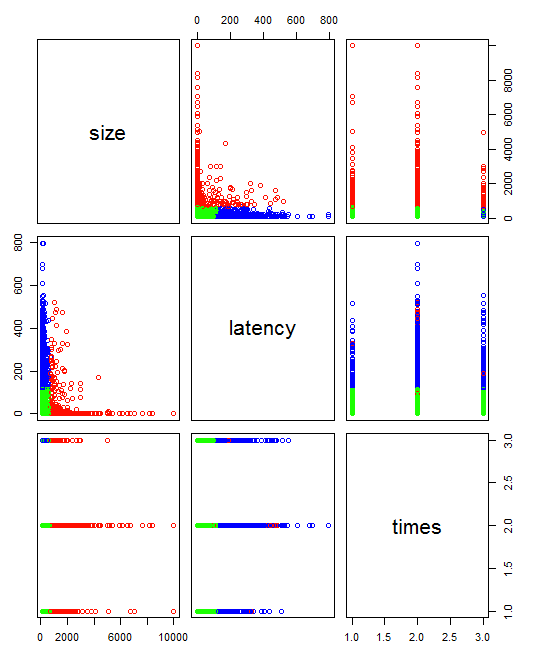
where B(K) is between-cluster variation and W(K) is within-cluster variation. A higher CH value is desirable, because it indicates the clustering maximizes between-class variation while minimizes within-class variation.

K=2: the result converges and CH is 69779.33.



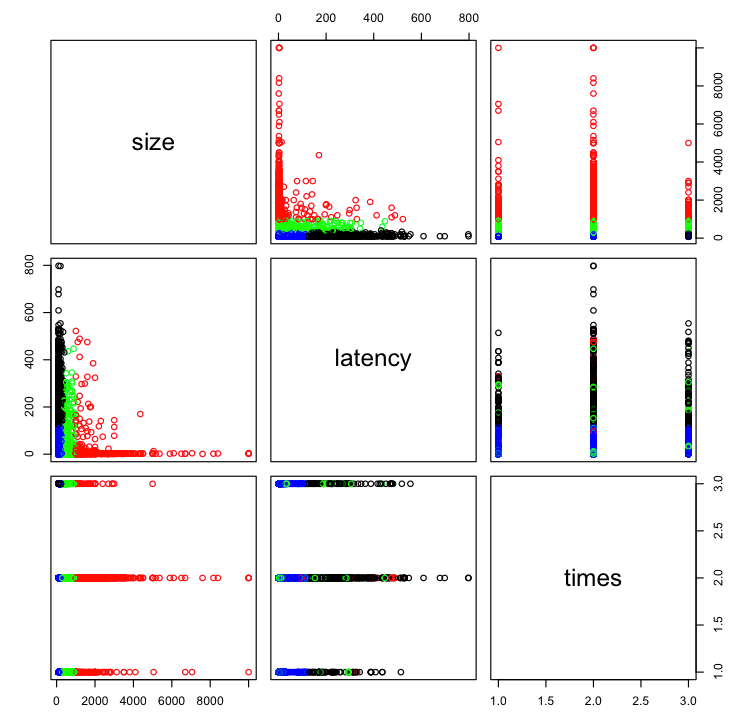
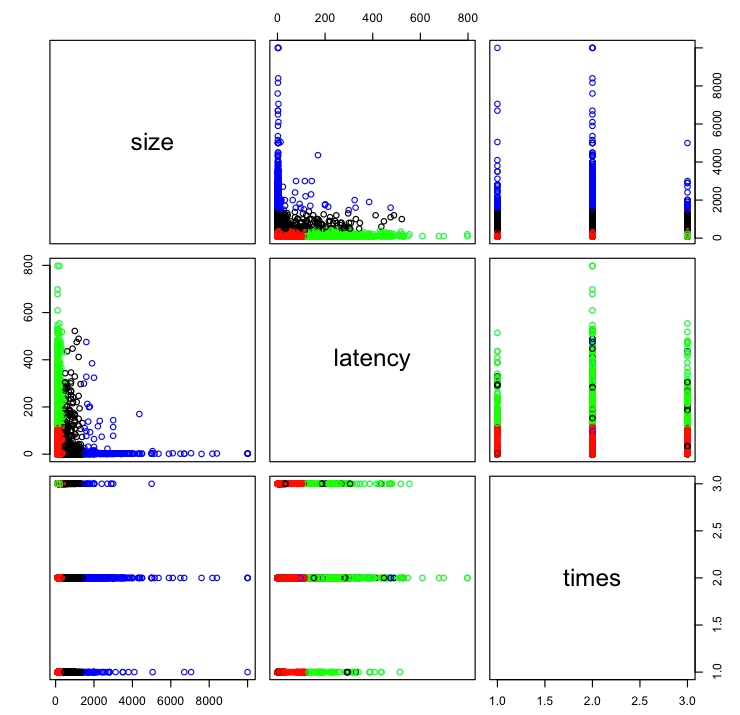
*Figure 1: k-means clustering with k=2*

K=3:

If K>2, we start to get different results in the end. This is common because k-means clustering does not ensure one single result. Different clustering will provide different CH value, and we choose the one with higher CH value.

*Figure 2: k-means clustering with k=3 (CH=47970.17)*

*Figure 3: k-means clustering with k=3 (CH=*85656.81*)*



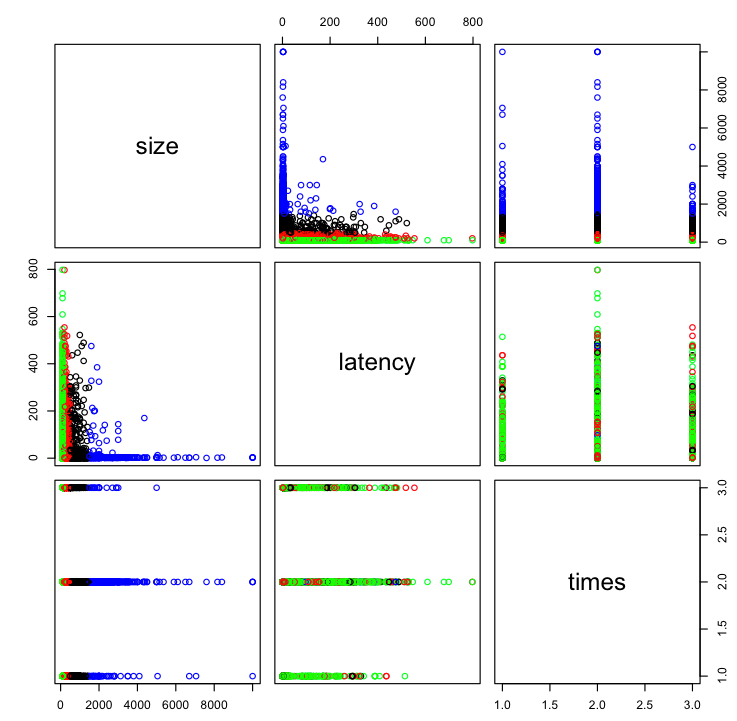
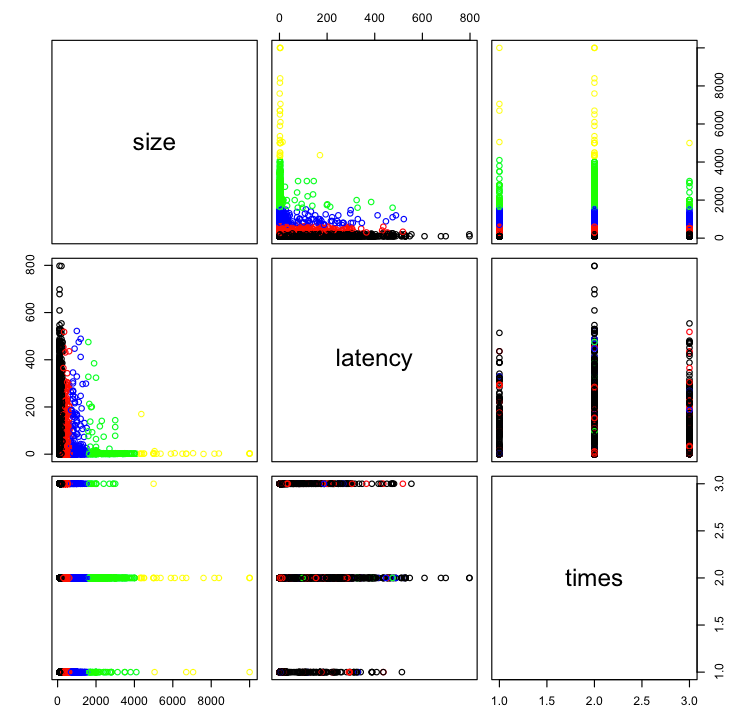


Figure 4: *k-means clustering with k=4 (CH=*66520.76*)*

Figure 5: *k-means clustering with k=4 (CH=*88403.78*)*

Figure 6: *k-means clustering with k=4 (CH=*113696*)*



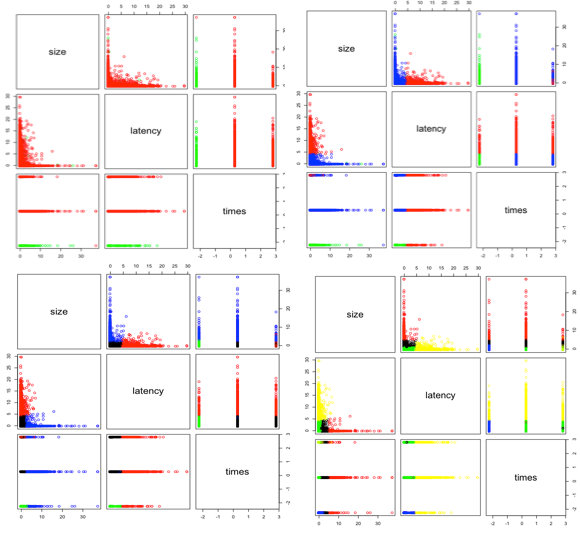
*Figure 7: k-means clustering with k=5 (CH=179403.9)*

Conclusion: from the plots, it seems that size has played a dominant role in the clustering. However, because size has the largest variance, it may distort the clustering result. Therefore, we try to standardize data before clustering.

**Scaling data**

The aim of clustering is trying to maximize between-class variation and minimize within-class variation, so standardizing variables is essential for clustering, otherwise some variable with large variance (like size in this case) will dominate the clustering.

**Centering and scaling: Xstd=**

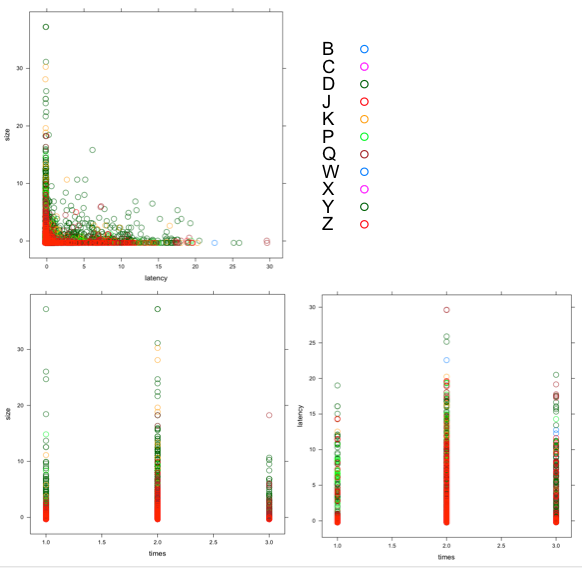
****

*Figure 8: k-means clustering with k=2 (CH=19538.14***)**

*Figure 9: k-means clustering with k=3 (CH=39115.74)*

*Figure 10: k-means clustering with k=4 (CH=65663.61)*

*Figure 11: k-means clustering with k=5 (CH=72427.49)*

In figure 8-11, we use standardized data to do clustering, with k=2-5. There is no single result, so we choose the clustering plot with highest CH value for each k. The patterns are quite different from unstandardized data, where size is no longer dominant. When k=2, we can see the clusters are merely depends on time. As k increases, latency becomes significant. Generally, those trades with high latency are classified in one group, regardless of trading time; early transactions with low latency are classified in one group, and low-latency transactions happening in midday and in the last ten minutes are classified in one group.