**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.

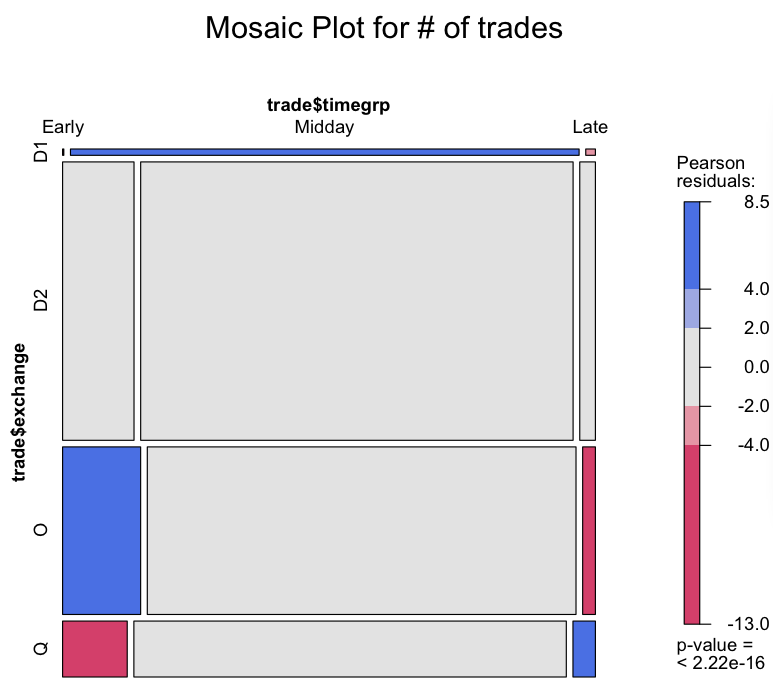


Figure 1

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

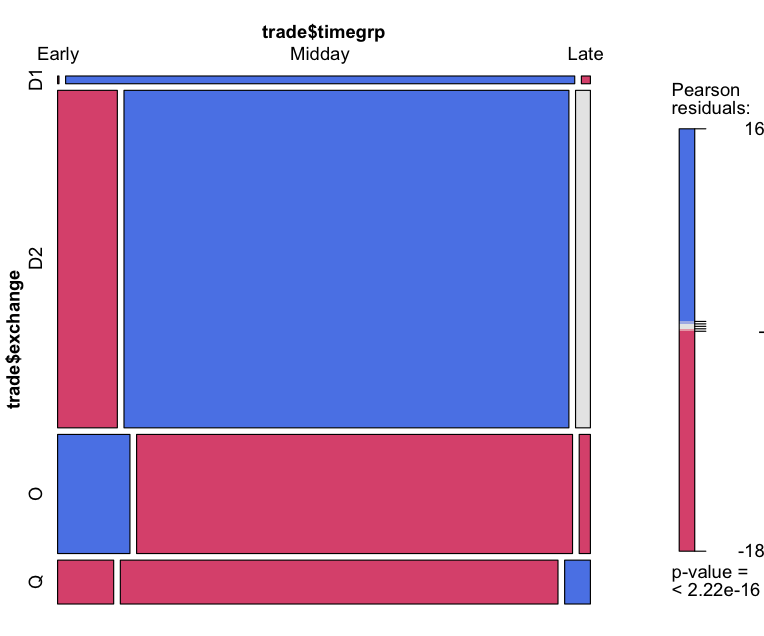


Figure 2

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)*

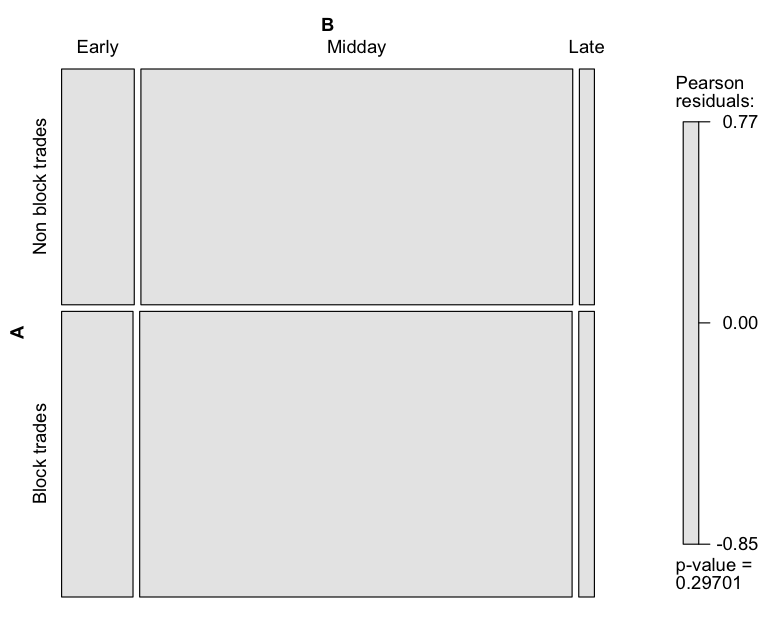


Figure 3

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

*Table 2 (Contingency Analysis for # of shares)*

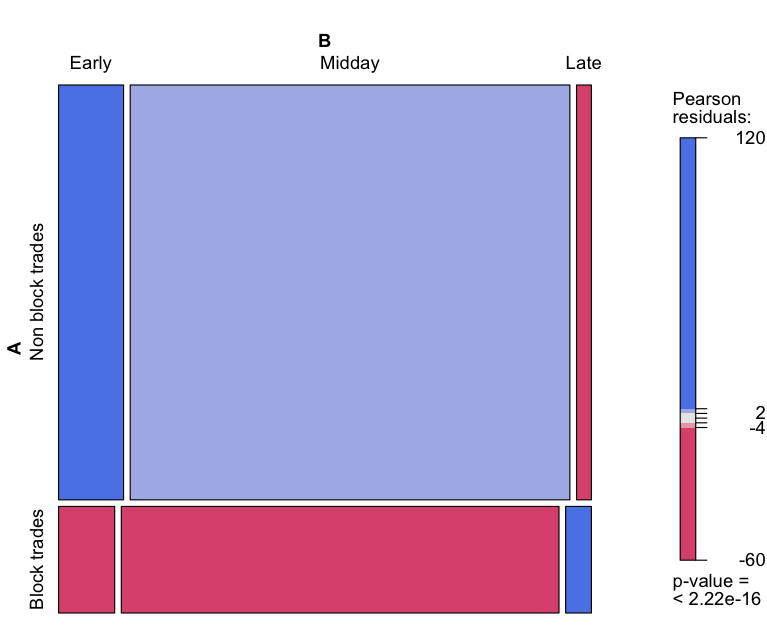


Figure 4

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**

**Model: In = ++**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Intercept | IMidday | ILate |
| 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:

**In = 7.866580 -4.243032IMidday -3.597384 ILate**

* 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.

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:**

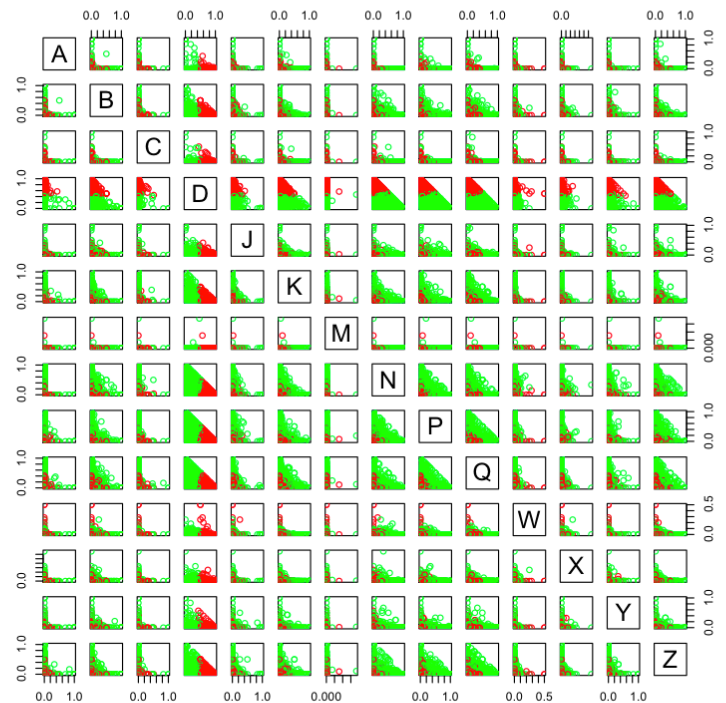


Figure 5: Clustering plot of early trading and k=2

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**

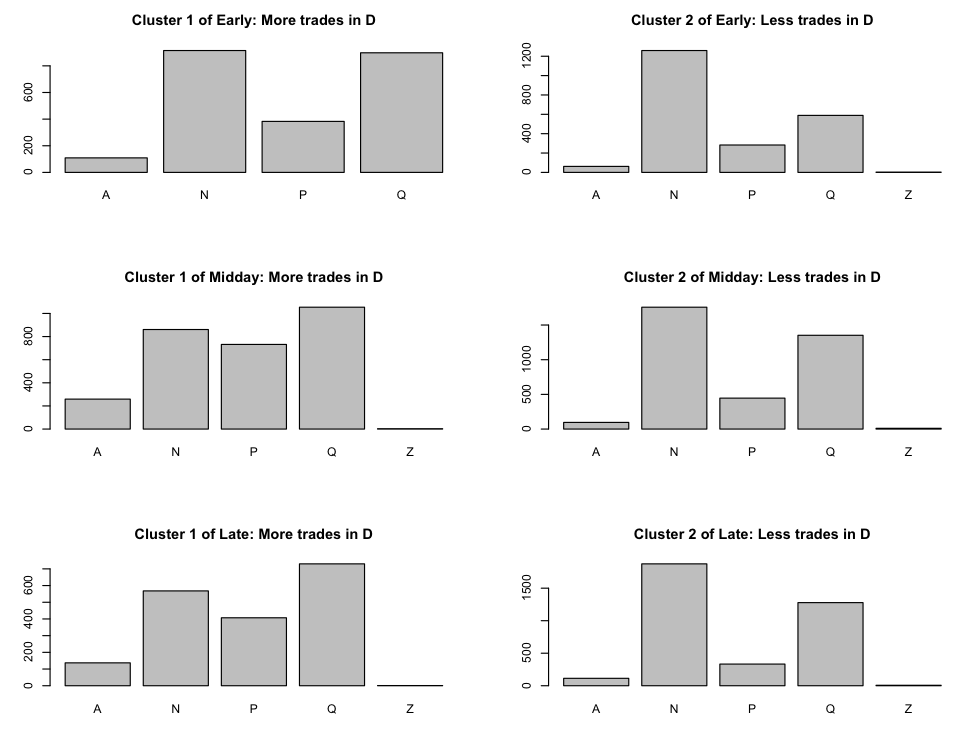


Figure 6: Barplots of Stocks’ Primary Exchange in different clusters

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.

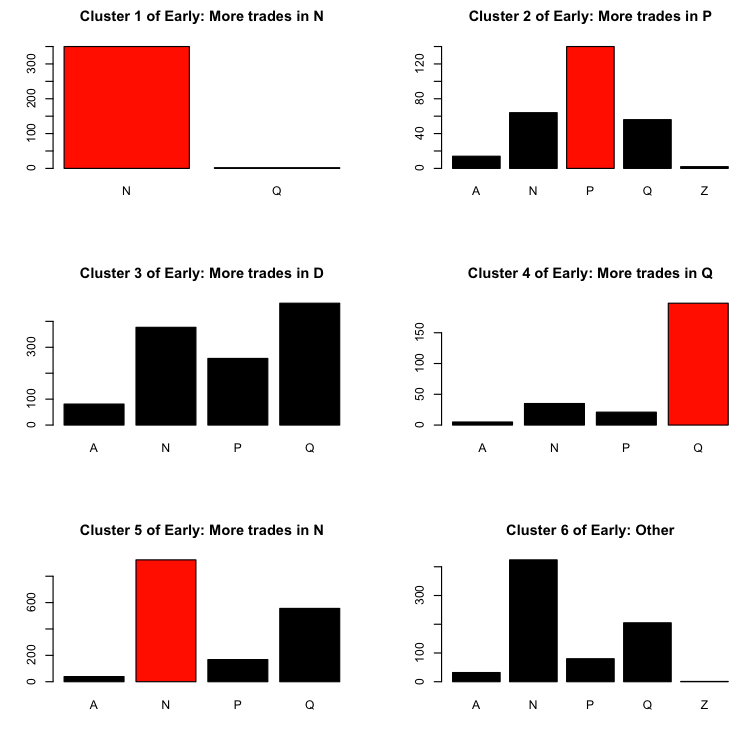


Figure 7: Barplots of Stocks’ Primary Exchange in 6 different clusters in early

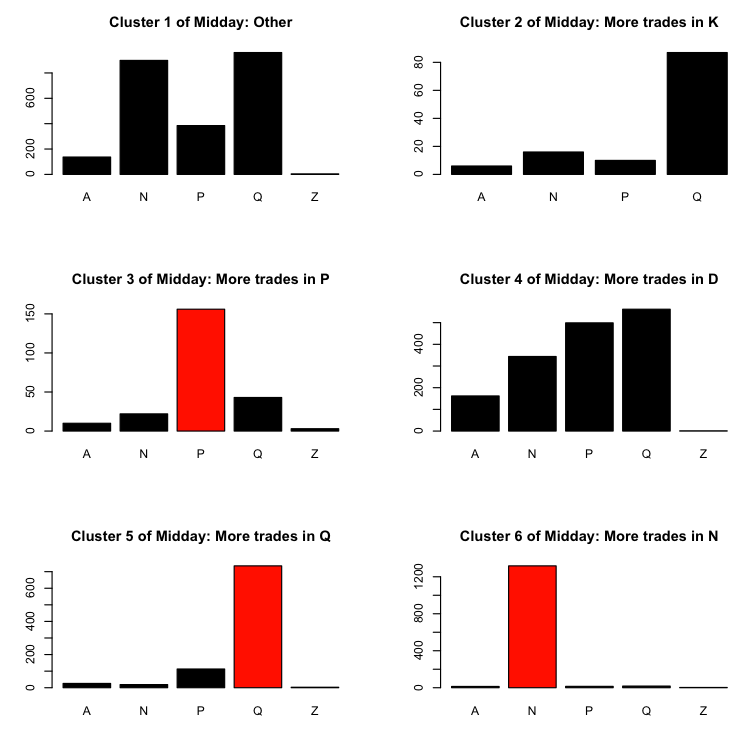


Figure 8: Barplots of Stocks’ Primary Exchange in 6 different clusters in midday

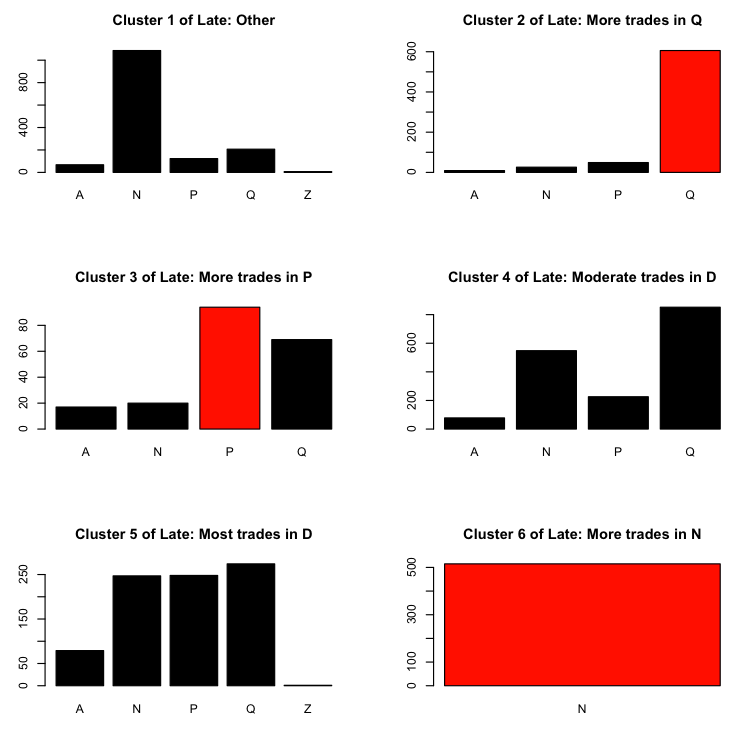


Figure 9: Barplots of Stocks’ Primary Exchange in 6 different clusters in late

**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.

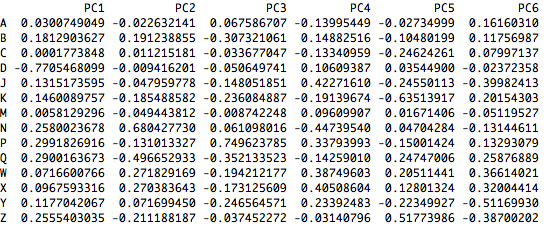


Table 9 (First 6 PC direction)

* Midday period



Table 10 (deviation PC can explain)

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

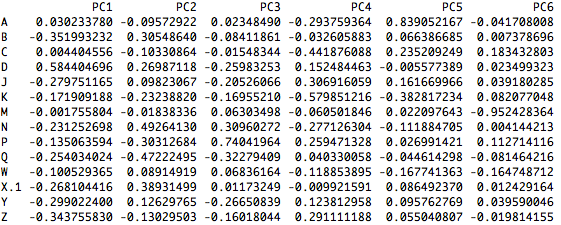


Table 11 (First 6 PC direction)

* Late period



Table 12 (deviation PC can explain)

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

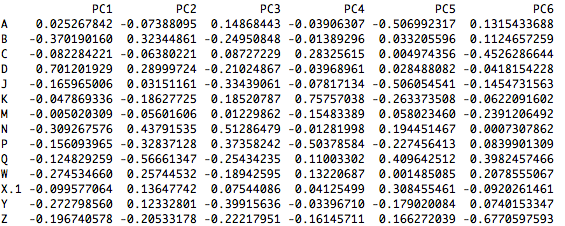


Table 13 (First 6 PC direction)