Final project 4415

Huilong An ha2399

Q1. [R] The dataset "evaluation.RData" consists of a total 5820 evaluation scores provided by students from an Ivy League university. The detailed description can be found in the file "README_eval.txt". The aim is to explore and describe the dependency pattern across different variables.

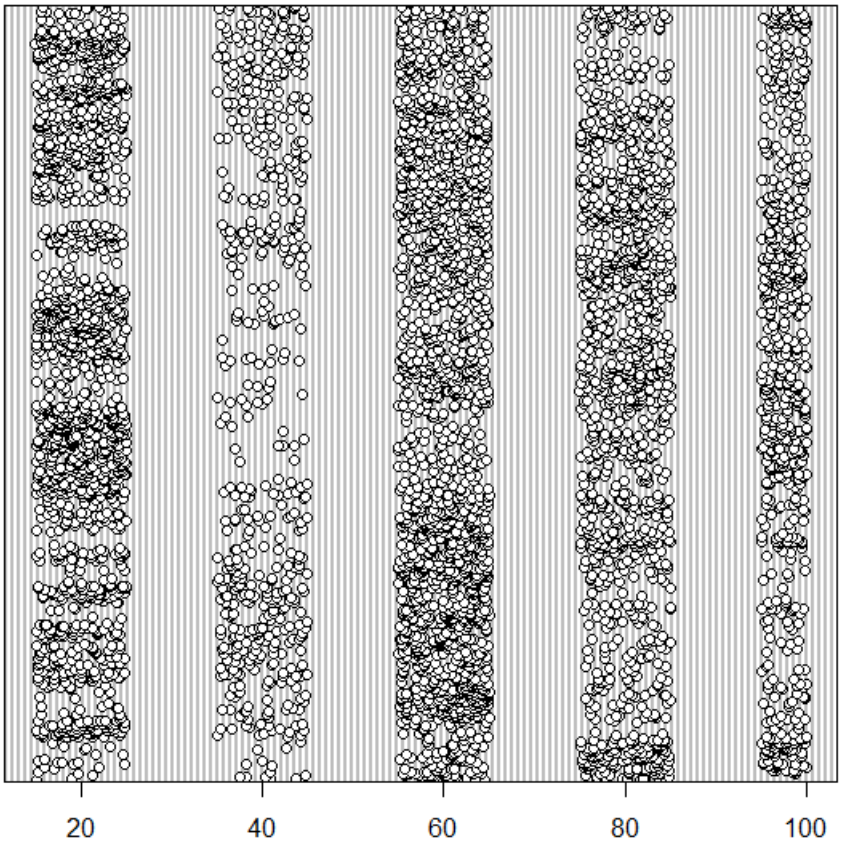
Step 1: taking a look at the data

Let us first take a look at the structure of the dataset:

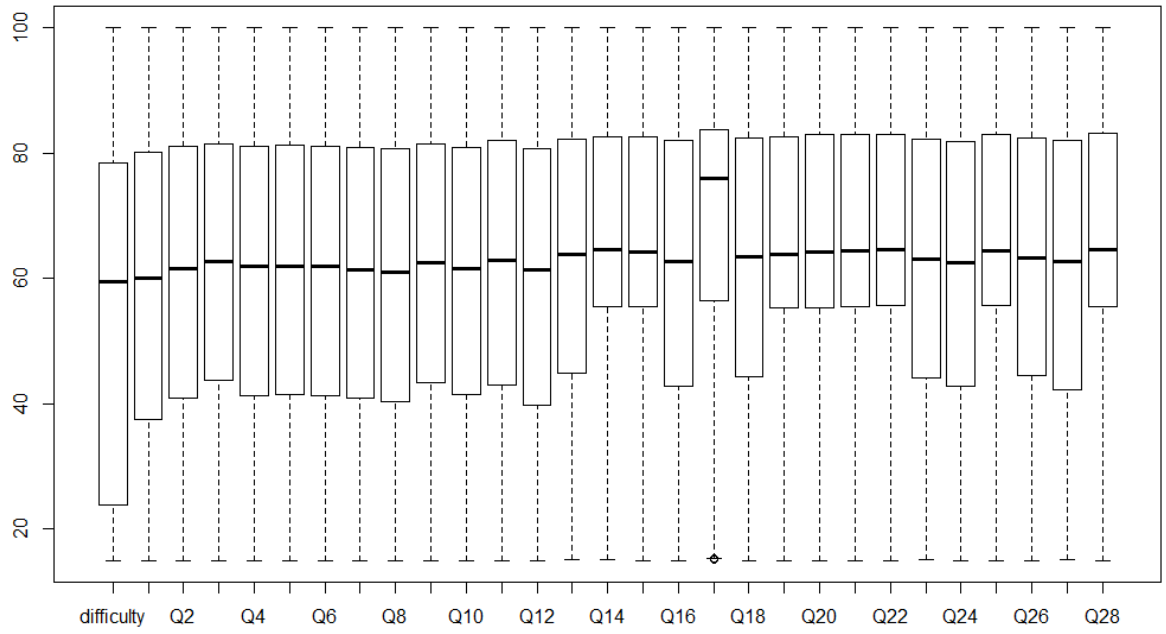
```
> str(score)
'data.frame':      5820 obs. of  34 variables:
 $ instr     : num  1 1 1 1 1 1 1 1 1 1 ...
 $ class     : num  2 2 2 2 2 2 2 2 2 2 ...
 $ nb.repeat : num  1 1 1 1 1 1 1 1 1 1 ...
 $ attendance: num  0 1 2 1 0 3 1 1 1 4 ...
 $ difficulty: num  81.6 59.9 78 60.7 15.1 ...
 $ Q1        : num  59.8 58.8 98.7 61.5 18 ...
 $ Q2        : num  60.2 59.5 98.7 62.2 17.5 ...
 $ Q3        : num  60.9 63.8 96.7 57.8 22.2 ...
 $ Q4        : num  55.4 60.5 97 62.8 24.6 ...
 $ Q5        : num  58.6 57.6 96.1 64.6 22.6 ...
 $ Q6        : num  58.7 56.9 97.1 60 19.7 ...
 $ Q7        : num  60.5 56.1 96.7 61.6 16.5 ...
 $ Q8        : num  56.3 63.9 97.4 56.7 24.2 ...
 $ Q9        : num  57.6 58.6 98.8 59.4 17.3 ...
 $ Q10       : num  58.7 56.5 98.8 57.9 17.2 ...
 $ Q11       : num  60.6 57.1 95.2 63.3 23.3 ...
 $ Q12       : num  64.5 57.6 97.9 63.8 19.7 ...
 $ Q13       : num  62.9 55.8 99.6 58.8 21.8 ...
 $ Q14       : num  63.7 58.7 98 60 22.6 ...
 $ Q15       : num  64.7 56.7 95.7 56 22.4 ...
 $ Q16       : num  62.4 55.1 96.6 56.1 22.6 ...
 $ Q17       : num  56.1 63.3 98.6 63.3 16.6 ...
 $ Q18       : num  55.2 61.8 96 62 18.5 ...
 $ Q19       : num  62.5 61 97.8 61.2 24.9 ...
 $ Q20       : num  56.5 61.5 96.4 58.2 15.5 ...
 $ Q21       : num  60 58.7 97.2 61.2 22.6 ...
 $ Q22       : num  60.4 61.5 96 57.2 20.4 ...
 $ Q23       : num  63 59.9 98.4 60 16.6 ...
 $ Q24       : num  55.7 61 99.8 61.5 24.9 ...
```

```
$ Q25       : num  57.5 55 96.9 55.1 20 ...
$ Q26       : num  62 61.6 97.4 62.5 15.1 ...
$ Q27       : num  56.7 59.6 97.4 56.9 24.3 ...
$ Q28       : num  59.2 63.3 95.1 62.8 19.9 ...
```

There are 33 variables and 5820 sample. For the first four variables, they are all qualitative variables with different levels. And if we draw a Cleveland dot chart, we can easily find from variable “difficulty” to “Q28”, there are five significantly well-separated intervals for their values, so we can use a category variable to well descript these variables if needed, taking “difficulty” as an example:



Then we are going to use boxplot to see if there are some extreme values in each question scores as an initial way to detect potential outliers:



All scores of every questions seems ordinary, expect that there exist extreme value in Q17, and it is the 5755th sample.

```
> which.min(score$Q17)
[1] 5755
```

Step 2: framework for analysis of dependencies

As we see in step 1, there are 4 category variables (instructor, class, repeat, attendance), and 29 semi-continuous variables (I mean, they can be well-coded as category variables). So, we consider to divide these 33 variables to two group, 4 category variables as a group and the other 29 variables as the other group. We will first detect the relationship among the 29 variables group.

For these 28 questions variables of these 29 variables, when we take a look at the questions in the description documents, we can find they can be treated as two types: the first type is all related to the course (the content, feedback, practicability etc.), question of this type ranges from 1st to 12th and the second type is all related to the instructor (clearness, expertise etc.), question of this type ranges from 13th to 28th. And when we draw the correlation plot of these 28 variables, we can find some blocks. So, here are two ways to detect the relationship between these 28 variables: canonical analysis and factor analysis. Here, I use the factor analysis, which gives a clear description of this relationship.

Once we get the reasonable result of factor analysis, we might consider to use factor scores to have good summary of the information contained these 28 variables. However, what we should do in this process is to provide evidence to show that these extracted factors indeed

can have a good summary of the data, and the factor scores are reasonable and well-interpreted. And for the difficulty variables, we might, intuitively, believe there is a causal effect between these 28 question variables and difficulty, so they might be strongly dependent. We will use some way to find if the intuition is right.

Then we will discover the dependency among 4 qualitative variables, using Person's Chi-square test.

The last step is to uncover the dependency between two variables group. Noticed that the effect of instructors and classes in 28 question scores might be contained in the information of these 28 question scores, so these four category variables could have a strong dependent relation with these 28 question variables. Here, we will mainly use some visualization methods.

Summary:

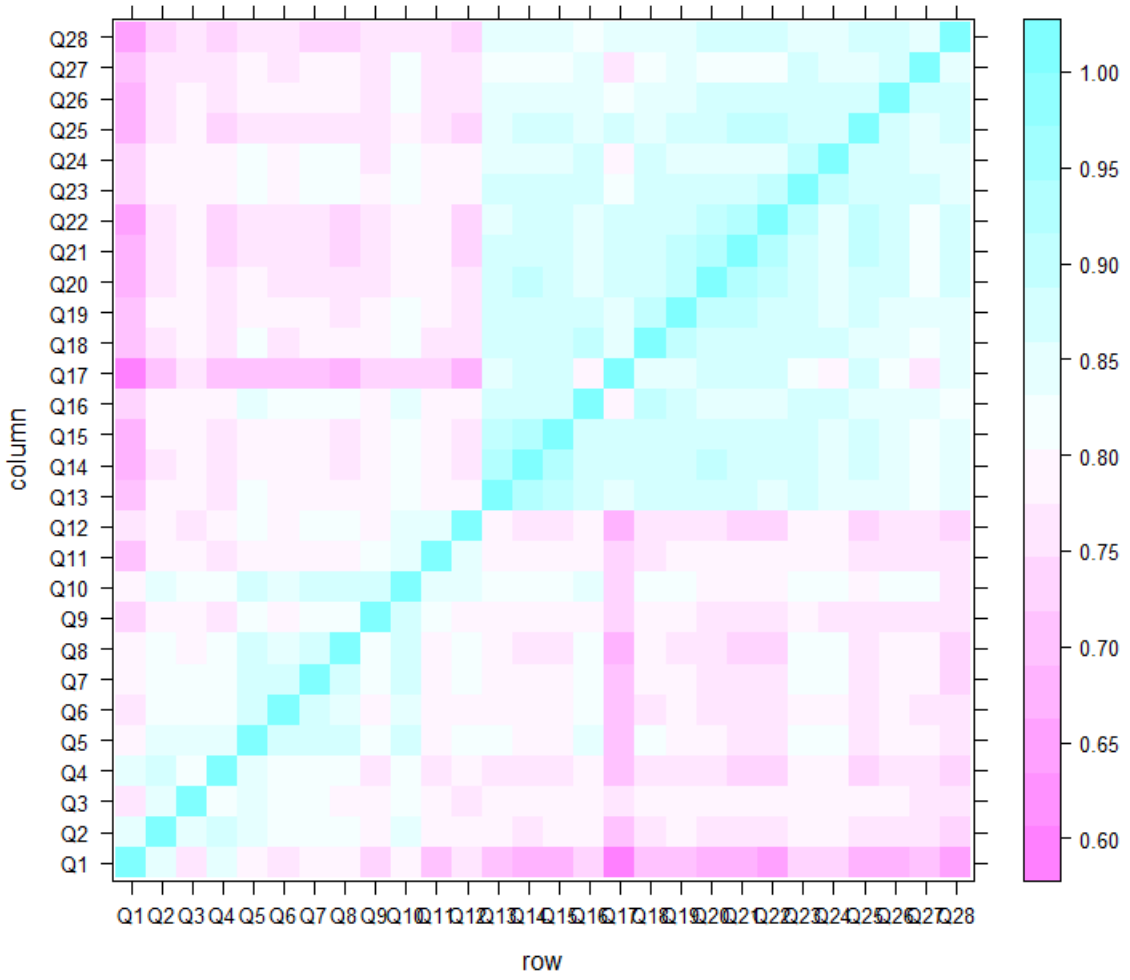
We separate 33 variables into tow part, the category variable part and evaluation variable part, and detect the relationship between these two part and within each part.

(1) Use factor analysis to extract reasonable number of factors of theses 28 question variables, and prove its appropriateness, then to find if the scores in each factors are a good summary of data in some sense. Also find the relationships among these 28 variables.
(2) Descript and detect the relationship between these 28 question and difficulty score. And find the relationship between difficulty and factor scores.
(3) Use Pearson's test to detect the relationships within 4 category variables, and draw some visualization plot to show these relationships.
(4) Use ANOVA methods to find the relationship between these two part.

Step2: Factor Analysis

2.1 Within 28 question variables group

We first take a look at the correlation plot of these 28 question variables.



They are all highly and positively correlated to each other, and we know factor analysis works well in a high-correlated dataset. As what indicates in this plot, there might exist some "blocks" in these 28 variables.

```
> faa=factanal(ques,factors=2,method="mle",rotation="varimax",scores="regression")
> faa$loadings

Loadings:
    Factor1 Factor2
Q1  0.372   0.777
Q2  0.490   0.763
Q3  0.564   0.682
Q4  0.471   0.765
Q5  0.501   0.788
Q6  0.496   0.768
Q7  0.461   0.813
Q8  0.457   0.806
```

```
Q9  0.541   0.695
Q10 0.522   0.784
Q11 0.563   0.676
Q12 0.483   0.745
Q13 0.752   0.550
Q14 0.793   0.509
Q15 0.789   0.509
Q16 0.705   0.604
Q17 0.824   0.389
Q18 0.758   0.535
Q19 0.787   0.514
Q20 0.819   0.474
Q21 0.839   0.446
Q22 0.840   0.444
Q23 0.751   0.561
Q24 0.708   0.589
Q25 0.821   0.461
Q26 0.747   0.537
Q27 0.690   0.566
Q28 0.807   0.450

                Factor1 Factor2
SS loadings      12.644  11.061
Proportion Var    0.452   0.395
Cumulative Var    0.452   0.847
```

We use a plot to visualize these two factors to have a good interpretation about these two factors:

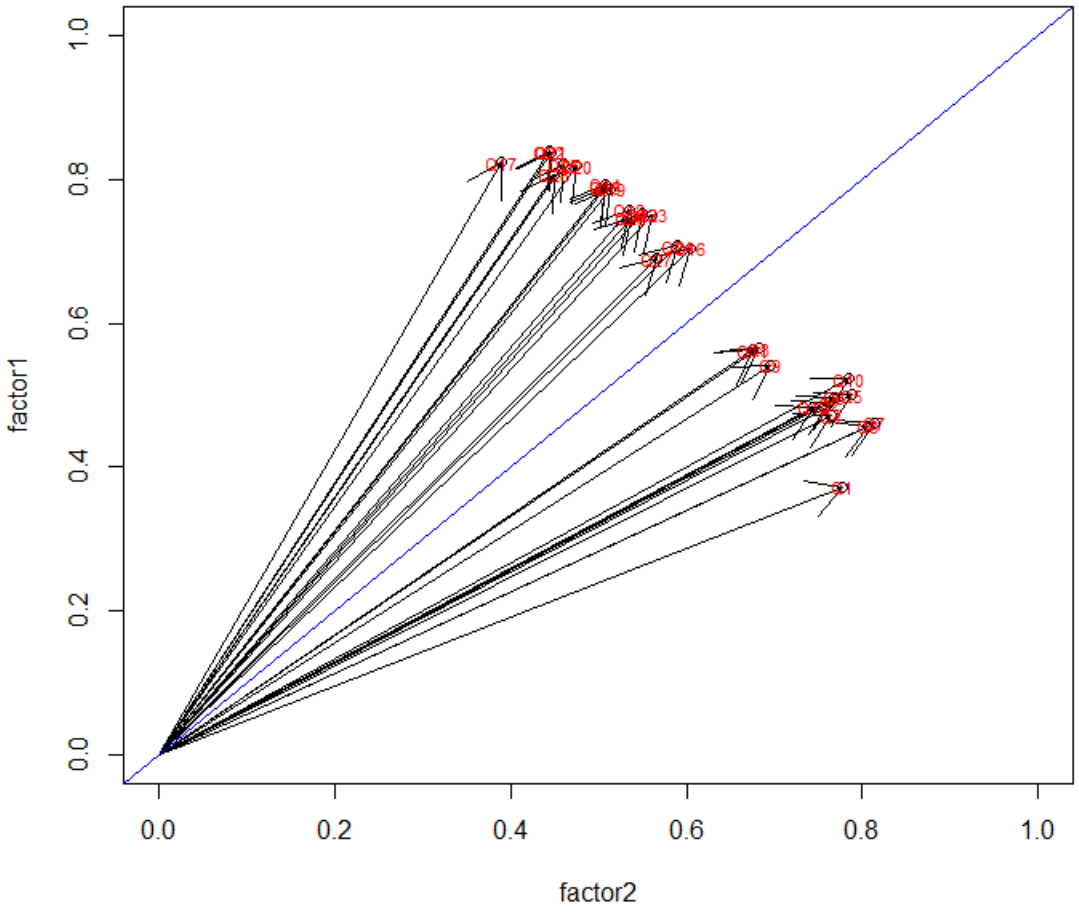
In the plot below, we can find the first factor can be labeled as “evaluation of instructor”, because we can find from the plot that variable “Q13” to “Q28” loads mainly on the first factor, and the second factor can be labeled as “evaluation of course”, as it shown in plot “Q1” to “Q12” loaded much on the first factor.

However, when we check the result of the test on the sufficiency of the number of factors:

```
> sapply(1:4,function(f) factanal(ques,factors=f,method="mle",rotation="varim
ax")$PVAL)
objective objective objective objective
        0         0         0         0
```

We find 2 factors are probably not sufficient (P-value is 0, which, in statistical significance test, can be interpreted as there is strong evidence to reject null hypothesis). But it is not necessary a case, because test hypothesis here is just a reference. We can check the goodness of correlation matrix recovered by these two factors compared to the initial

correlation matrix, because FA is such a method to collect the information content in the second moment of the variables.



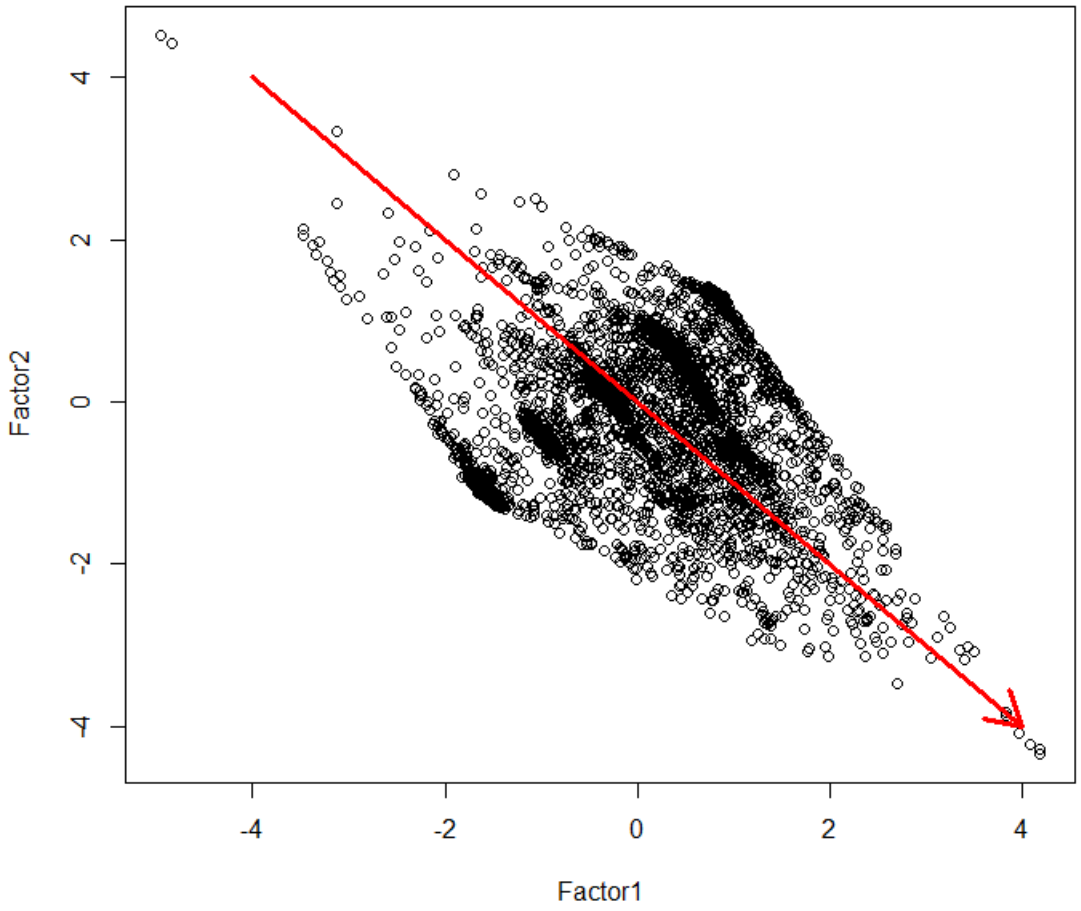
Check on predicted correlation matrix by two factor:

```
> #tcrossprod+diag term is the predicted correlation matrix and cor() is the
observed cor mat
> max(abs(round(tcrossprod(faa$loadings)+diag(faa$uniquenesses)-cor(ques),
3)))
[1] 0.08
> #min correlation in the observed cor mat
> min(cor(ques))
[1] 0.6053121
```

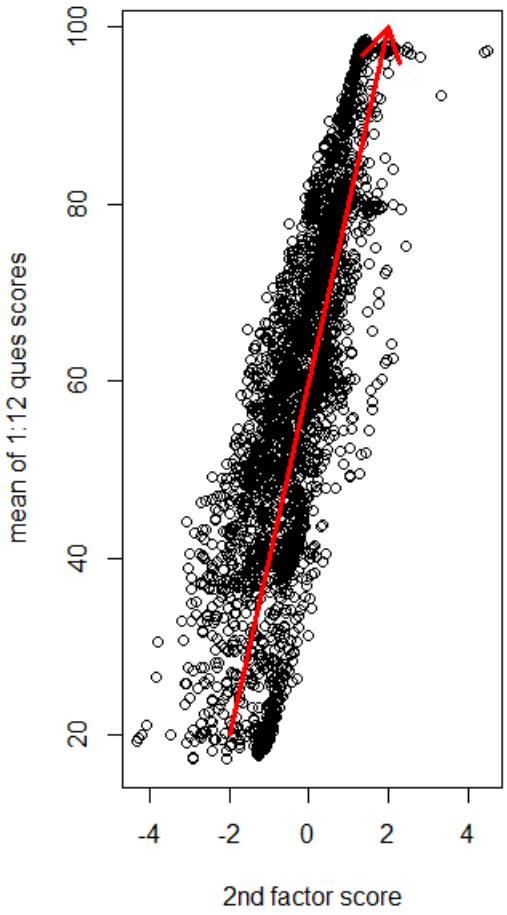
We can find the maximum difference between predicted correlation matrix and observed one is 0.08. Due to the high correlation among these variables, 0.08 is acceptable.

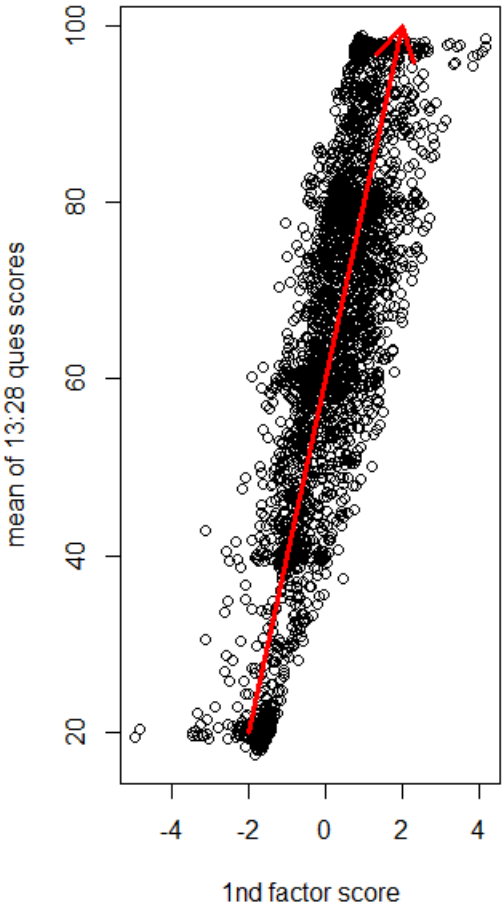
Estimated factor scores and interpretation:

We use the “regression” method to estimate factor scores, and get the plot:

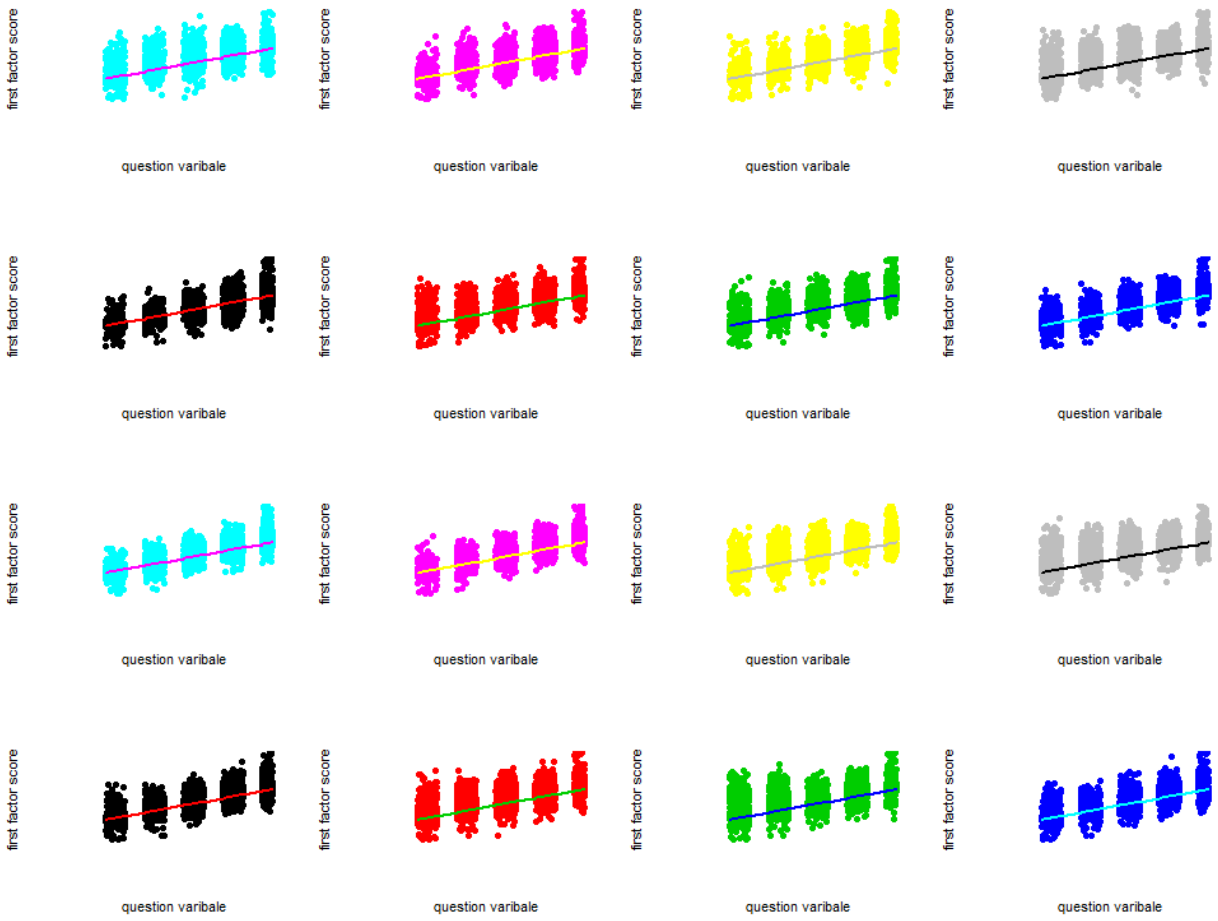


The trend in this plot is obvious, factor 1 and factor 2 are negatively related. We first need some check to verify whether these estimated factor scores are a good summary of data. We use the mean of first 12 question variables and that of 13 to 28 versus second factor score and first factor score, then we get the plot below.





And the plot below is the plot of first factor score versus each question variable, and the lines in the plot are the Lowess smoother, which gives us an indication of the relationship between two variables. We can find such relationships are all linear. (The second factor versus each of 1:12 question variables plot also shows linear relationship)



We can find that the interpretation of the factor scores is reasonable in some sense: a higher estimated factor scores represents a higher evaluation. So, we can use these factor scores to have a good summary of data of these 28 variables.

We can also pick out a point to see:

```
> identify(scores_f,col.index="red")
[1] 293
> scores_f[293,]# scores_f is the estimated factor score
  Factor1   Factor2
 1.855918 -0.882597
> ques[293,]
          Q1       Q2       Q3       Q4       Q5       Q6       Q7       Q8
     Q9       Q10       Q11       Q12
293 58.31028 59.82883 97.94737 59.61836 99.98711 24.47921 55.13886 96.47558 7
9.31888 57.29313 97.57215 20.44811
          Q13       Q14       Q15       Q16       Q17       Q18       Q19       Q20
    Q21       Q22     Q23       Q24
```

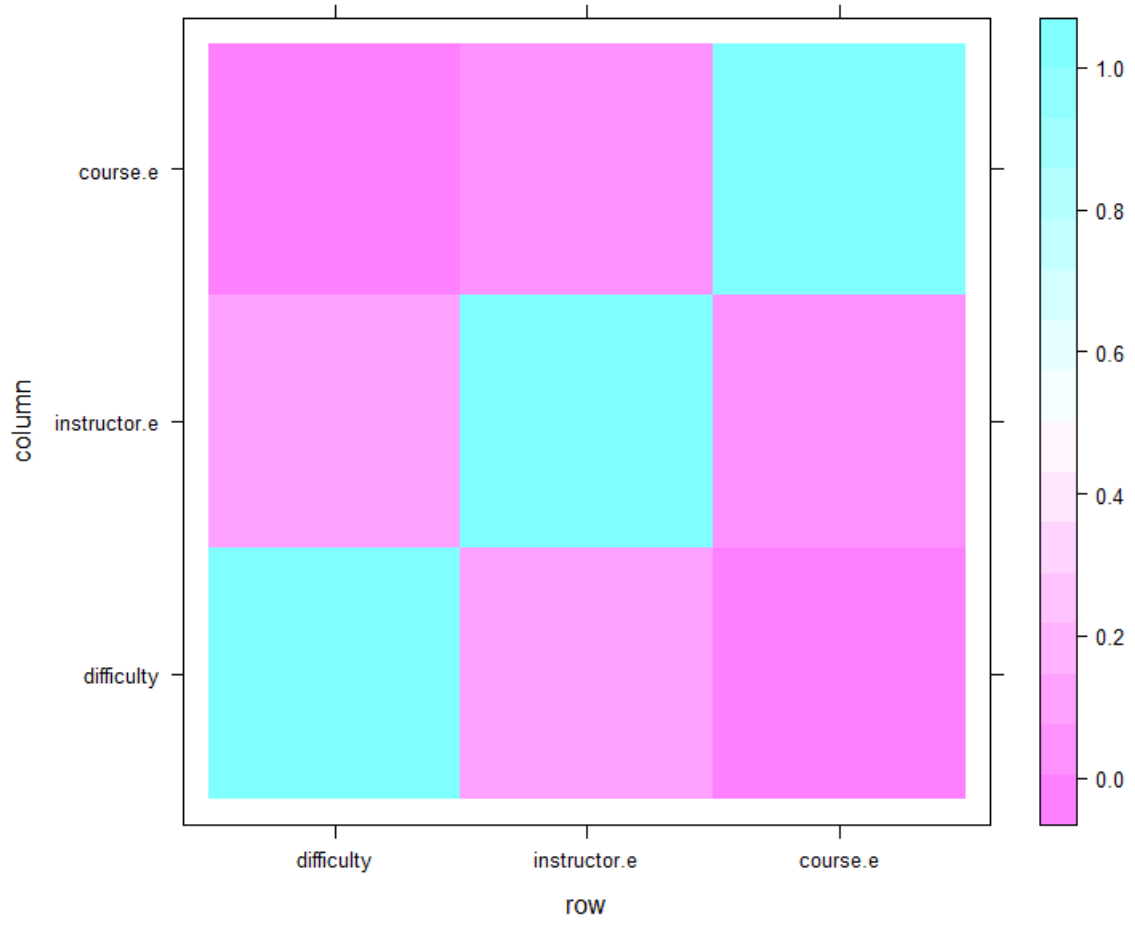
```
293 98.72495 98.91618 98.75346 55.26009 95.84988 95.95024 63.73028 99.44656 9
8.51387 99.87134 96.708 96.56825
         Q25      Q26      Q27   Q28
293 96.68909 62.70154 63.60167 97.18
```

The first 12 question scores are overall low, and the question scores from 12 to 28 are very high.

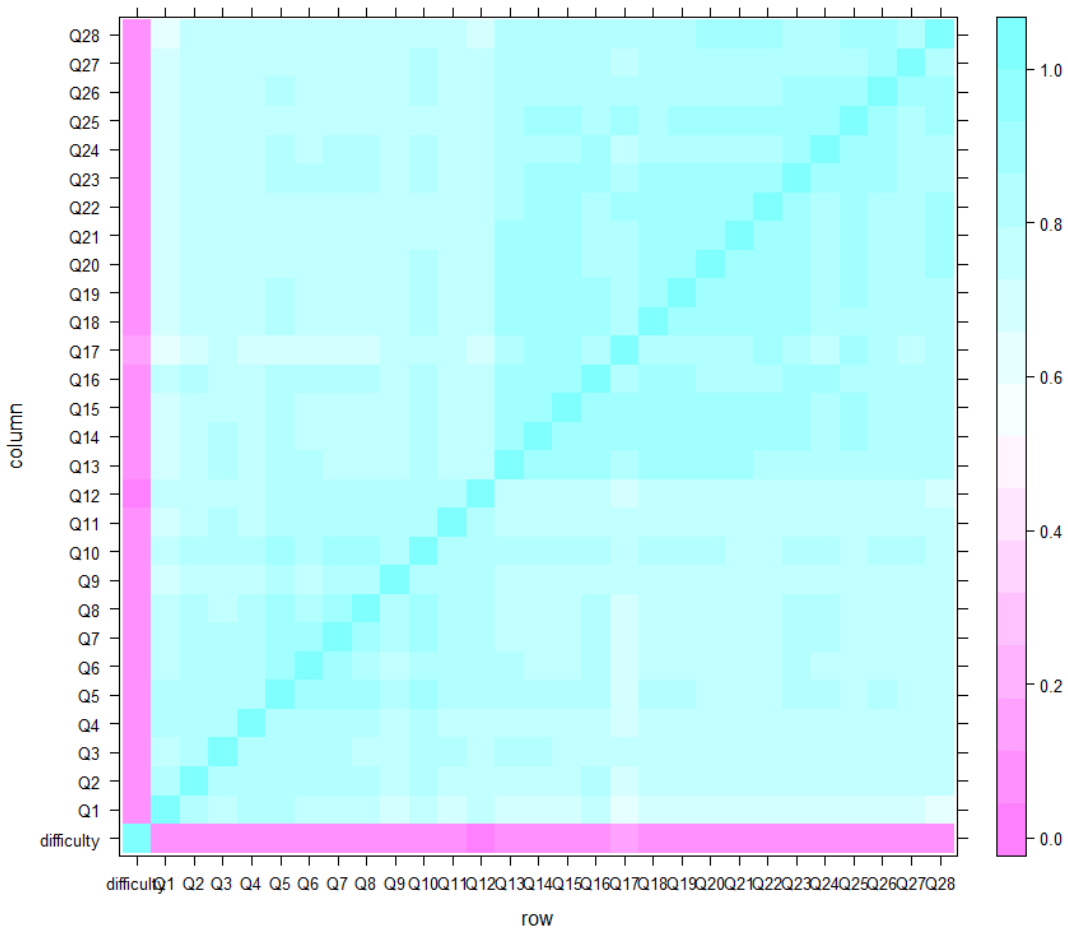
Then, we look back to the 1st factor score versus 2nd factor score plot. The trend in this plot can be interpreted as: students who have a higher evaluation on "course" tend to have a lower evaluation on the "instructor". This is the dependency between these two groups of 28 question variables.

2.2 the relationship between difficulty and 28 questions

The correlation plot shows, the correlation between difficulty and these 28 questions are very weak.

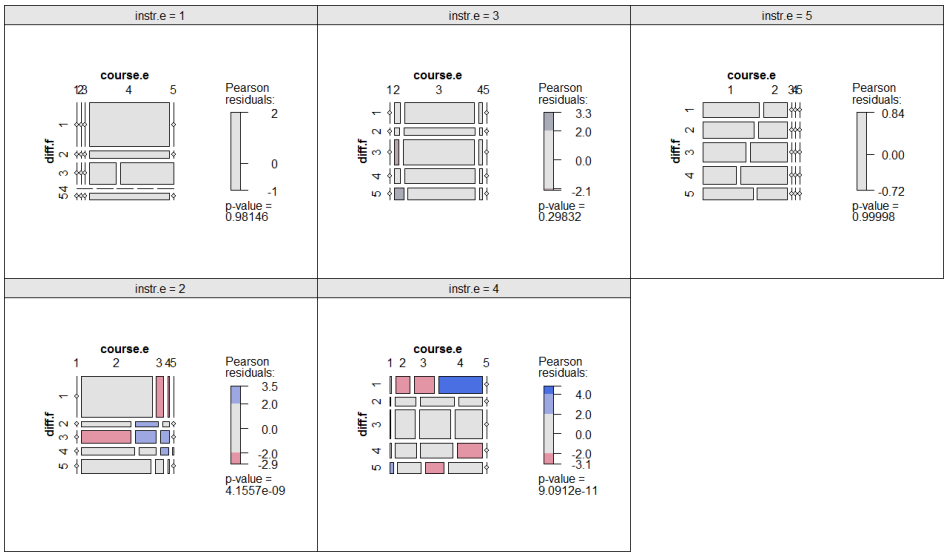


Just the same result as the correlation plot of these 29 variables:



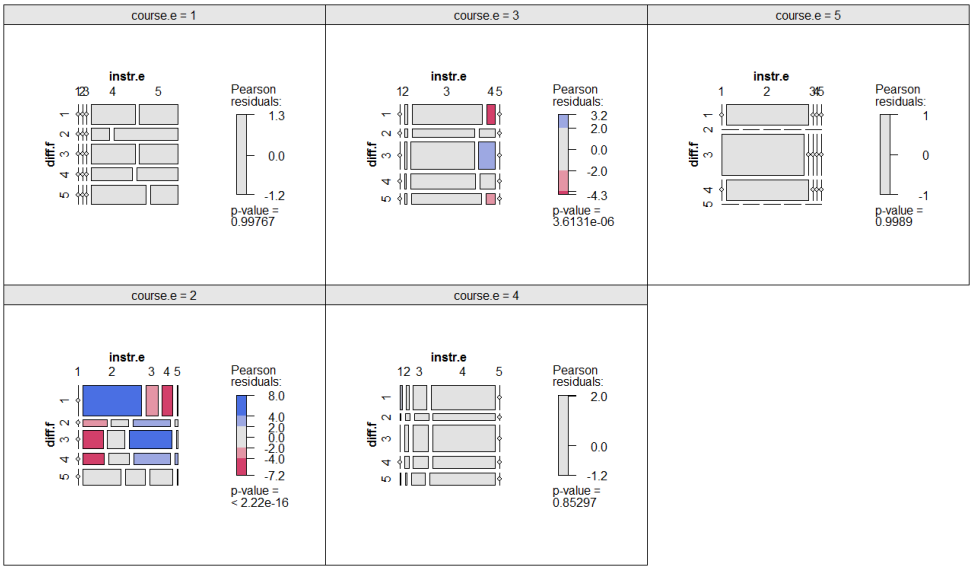
It seems there is no much dependency between the evaluation questions and difficulty, which is not as I expected.

Then, we turn the factor scores into category variables in 5 levels, which represents the degree of the evaluations (this is because all the 28 question variables can be turned into category variables in 5 levels). And we use the Pearson's chi-square test to see if we can find something.

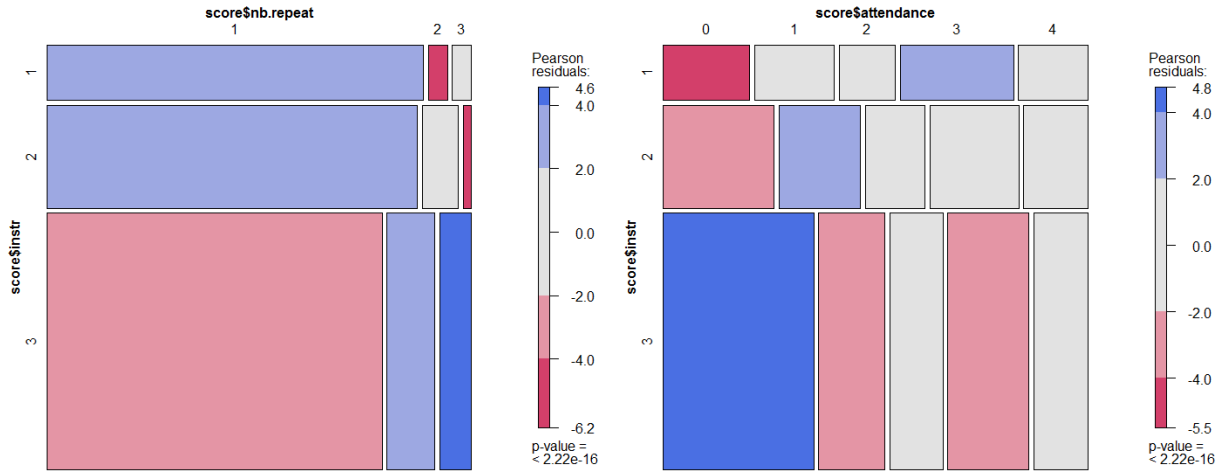


We first perform Pearson's test marginally on the degree of "evaluation on instructor", then we find in the 2nd and 4th degree of evaluation on instructor, the difficulty seems to be dependent with evaluation on course, that is to say, in medium low and medium high evaluation on instructor group, the difficulty is dependent with evaluation on course.

Then we perform Pearson's test marginally on the degree of "evaluation on course", then we find in 2nd and 3rd degree of evaluation on instructor, the difficulty seems to be dependent with evaluation on instructor. In other words, in medium low and medium evaluation on course group, the difficulty is dependent with evaluation on instructor.



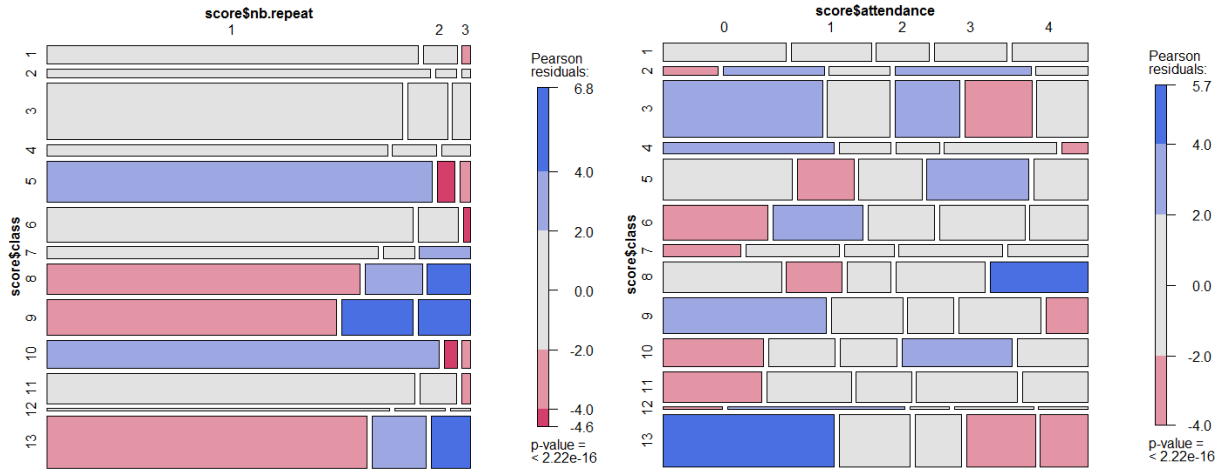
Step3: Dependency within the category variables group



Intructor versus nb.repeat: (right)

We can find instructor and nb.repeart are not independent, and the thrid instructor seem to have high nb.repeat value.

And instrctor and attendance are also not independent, the first and second instructor seems have high attendance.

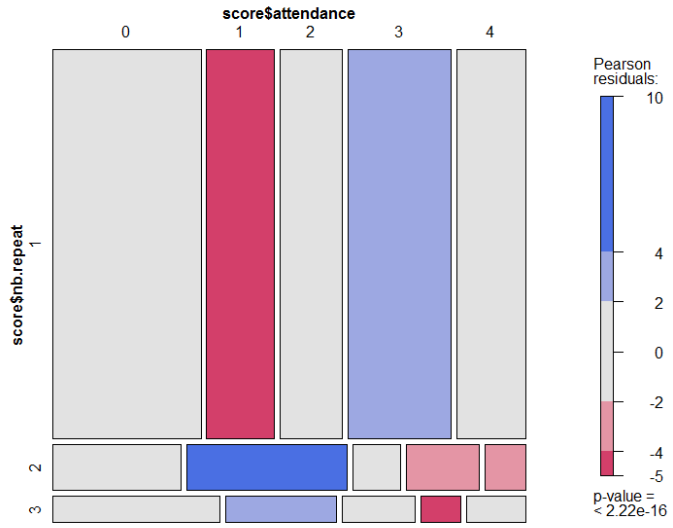


nb.repeat versus class and class versus attendance

We can find that class is not independent of nb.repeat and attendance. We can find 9th class have high nb.repeat value. And the attendance is also related to the calss.

nb.repeat versus attendance

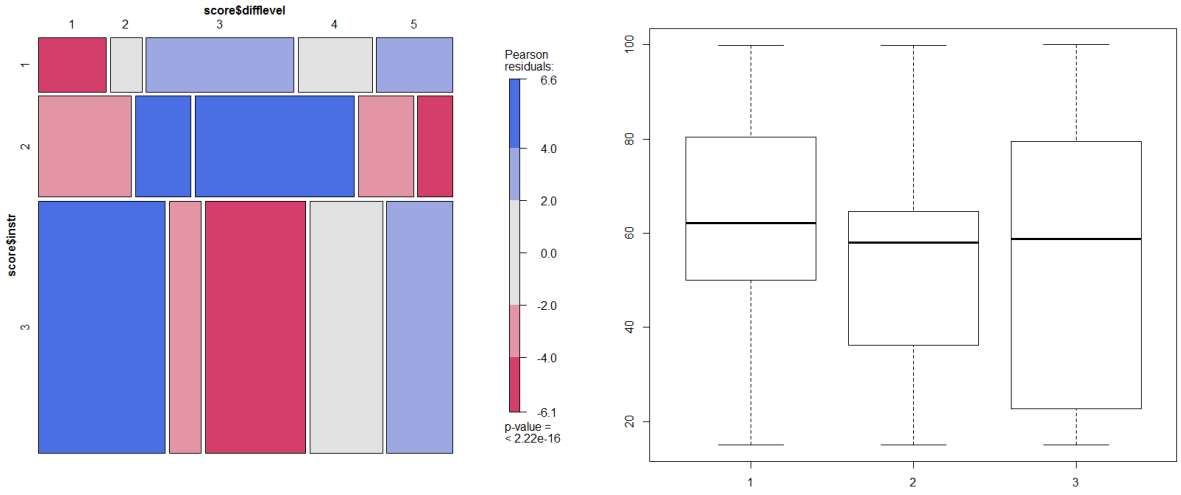
They are also not independent, nb.repeat is more likely to associated with a low attendence value.



Step 3: Dependency between two variable groups

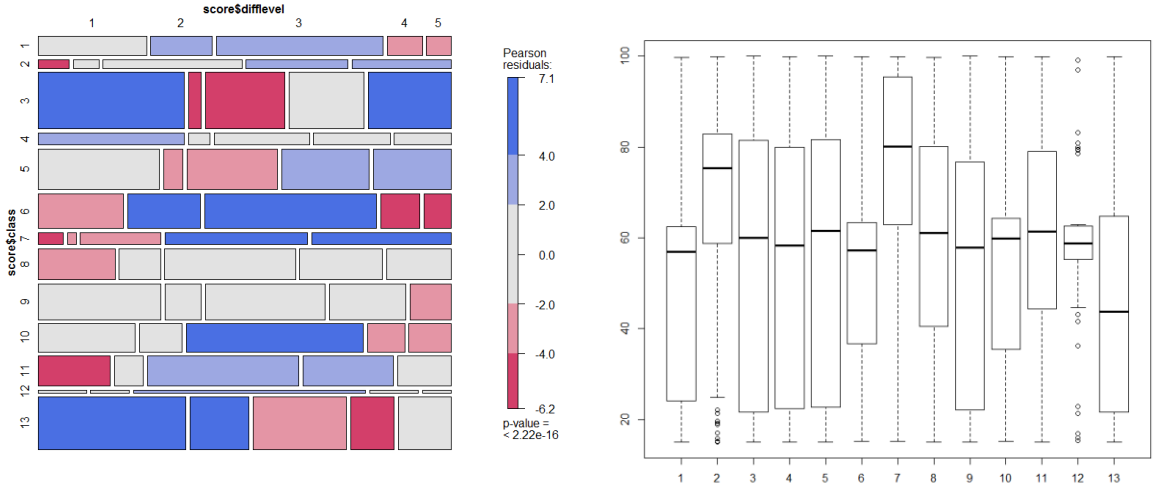
We only show two pairs of variable we are interested in:

Instructor versus difficulty:



Here, I divide difficulty to 5 different levels, and use Chi-square test, which show they are not independent, and instructor 1 and 3 seems have higher difficulty values than instructor 2.

Class versus difficulty:



We can find class is also not independent from difficulty, and the thrid and seventh class seems to be most difficult for students.

Then we want to have a look at the dependency between “evaluation on instructor” (group of 16 variables) and the qualitative variables. Here, we use ANOVA’s thought. We will extract class 10 first, as it only taught by instructor 2, which will cause some computation problem.

Here, we use the model:

$$y = \mu + \alpha + \beta + \gamma + \delta + \varepsilon$$

with global test:

$$H_0: \alpha = \beta = \gamma = \delta = 0$$

$$H_1: \text{at least one of } \mu, \alpha, \beta, \gamma, \delta \text{ is not } 0$$

Here, $\mu, \alpha, \beta, \gamma, \delta$ represent mean, effec of instructor, effect of class, effect of nb.repeat, and effect of attendance. And ε represents random error term.

```
Call:
lm(formula = scores_f[, 1] ~ as.factor(score$instr) + as.factor(score$class)
+
    as.factor(score$nb.repeat) + as.factor(score$attendance))

Residuals:
    Min      1Q  Median      3Q     Max
-5.0703 -0.5956  0.0548  0.6267  4.6364

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -0.33586    0.04855  -6.917 5.09e-12 ***
as.factor(score$instr)2      0.21013    0.06969   3.015 0.002580 **
as.factor(score$instr)3     -0.34659    0.12239  -2.832 0.004644 **
as.factor(score$class)2      0.16337    0.09070   1.801 0.071710 .
as.factor(score$class)3      0.29594    0.11810   2.506 0.012244 *
as.factor(score$class)4      0.52962    0.13299   3.982 6.90e-05 ***
as.factor(score$class)5      0.41970    0.11981   3.503 0.000463 ***
as.factor(score$class)6     -0.01021    0.06682  -0.153 0.878506
as.factor(score$class)7     -0.03375    0.08172  -0.413 0.679595
as.factor(score$class)8      0.79609    0.12133   6.562 5.79e-11 ***
as.factor(score$class)9      0.49562    0.12037   4.117 3.88e-05 ***
as.factor(score$class)11     0.02768    0.06859   0.404 0.686563
as.factor(score$class)12     0.32049    0.18527   1.730 0.083710 .
as.factor(score$class)13     0.23381    0.10848   2.155 0.031176 *
as.factor(score$nb.repeat)2 -0.09249    0.04231  -2.186 0.028833 *
as.factor(score$nb.repeat)3 -0.10042    0.05406  -1.858 0.063284 .
as.factor(score$attendance)1  0.18003    0.03697   4.870 1.14e-06 ***
as.factor(score$attendance)2  0.34066    0.03977   8.566  < 2e-16 ***
as.factor(score$attendance)3  0.47245    0.03448  13.702  < 2e-16 ***
as.factor(score$attendance)4  0.45260    0.03911  11.573  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9357 on 5800 degrees of freedom
Multiple R-squared:  0.07755,  Adjusted R-squared:  0.07453
F-statistic: 25.66 on 19 and 5800 DF,  p-value: < 2.2e-16
```

The global test(F test) shows we should reject null hypothesis. And the ANOVA table:

```
Analysis of Variance Table

Response: scores_f[, 1]
                             Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(score$instr)        2   41.9  20.965 23.9435 4.407e-11 ***
as.factor(score$class)       11  152.9  13.903 15.8781 < 2.2e-16 ***
as.factor(score$nb.repeat)    2   11.7   5.871  6.7056  0.001234 **
as.factor(score$attendance)   4  220.4  55.092 62.9188 < 2.2e-16 ***
Residuals                  5800 5078.5   0.876
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA table shows class and attendance have the most power to explain the variation in observations. We mainly foucs on the relationship between class as well as attendance and “evaluation on instructor” Although there exist some problems in this method, but the result is a good summary of such a relationship (when we try to plot some boxplot, we might find the relation described by parameter is really that in boxplots!):

Higher attendance will lead to a high evaluation on instructor (with same instructor and nb.repeat and class). And class 8 has a highest evaluation on instructor(with same attendance, instructor, nb.repeat).

Then we use “evaluation on class” as the response:

```
> anova(lm.fit)
Analysis of Variance Table

Response: scores_f[, 2]
                             Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(score$attendance)   4   18.1   4.525  5.0378 0.0004729 ***
as.factor(score$instr)        2  103.2  51.598 57.4475 < 2.2e-16 ***
as.factor(score$class)       11   83.3   7.575  8.4340 6.423e-15 ***
as.factor(score$nb.repeat)    2    3.9   1.969  2.1919 0.1117987
Residuals                  5800 5209.5   0.898
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can find instructor and class have the most power to explain variation in “evaluation on course”. And the corresponding parameters are:

```
as.factor(score$instr)2      -0.170160   0.070587  -2.411 0.015954 *
as.factor(score$instr)3      -0.204082   0.123960  -1.646 0.099745 .
as.factor(score$class)2       0.009628   0.091860  -0.105 0.916529
as.factor(score$class)3      -0.206128   0.119614  -1.723 0.084891 .
as.factor(score$class)4      -0.514405   0.134692  -3.819 0.000135 ***
as.factor(score$class)5      -0.124708   0.121344  -1.028 0.304122
```

```
as.factor(score$class)6      -0.038416   0.067675  -0.568 0.570286
as.factor(score$class)7      -0.418099   0.082762  -5.052 4.51e-07 ***
as.factor(score$class)8      -0.450776   0.122882  -3.668 0.000246 ***
as.factor(score$class)9      -0.186703   0.121914  -1.531 0.125717
as.factor(score$class)11     -0.143899   0.069473  -2.071 0.038376 *
as.factor(score$class)12     -0.340634   0.187642  -1.815 0.069523 .
as.factor(score$class)13     -0.328633   0.109868  -2.991 0.002791 **
```

We can find that instructor 1 seems receive higher “evaluation on course” under same conditions, and class 10 receives the highest “evaluation on course” under same conditions.

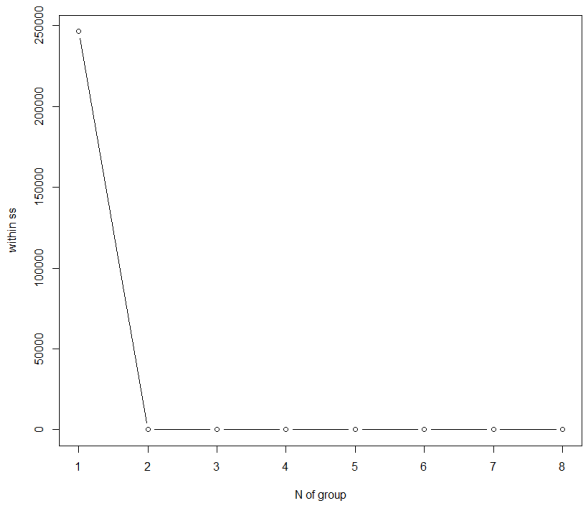
Problem 2:

Q2. [R] The dataset “Gesture.RData" composed by features extracted from 7 videos with people gesticulating. Each video is represented by two files: a raw file, which contains the position of hands, wrists, head and spine of the user in each frame; and a processed file, which contains velocity and acceleration of hands and wrists. See "README_gest.txt" for more detailed information on the dataset. The objective is to cluster each video frame (based on the extracted features) into different groups, according to the underlying gesture it comes from.

Solution:

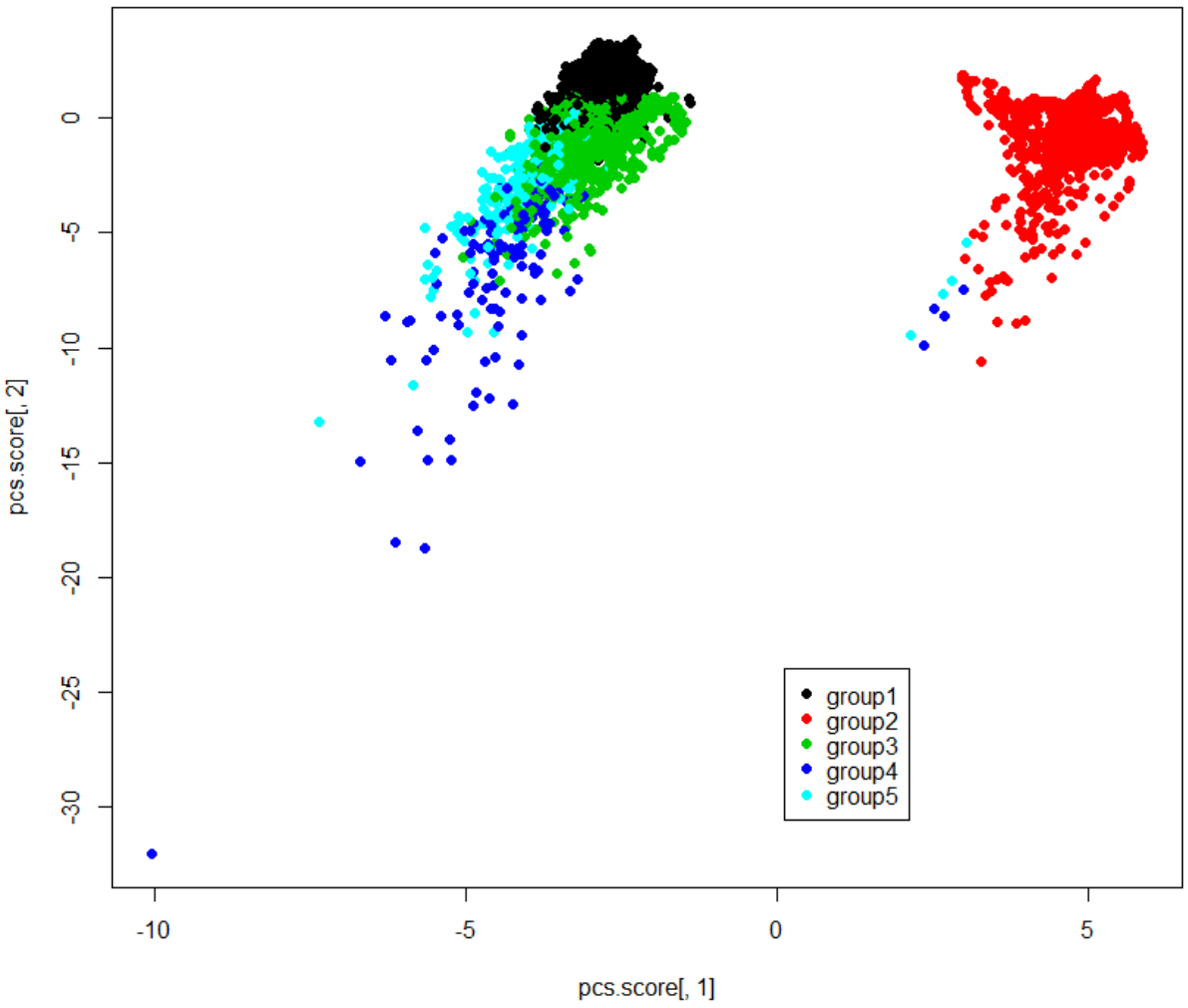
Here, I use K-means cluster method. The result is highly variable, so I run K-means cluster many times to get the acceptable result:

The within-group sum of squares plot is shown below: (after scaled)



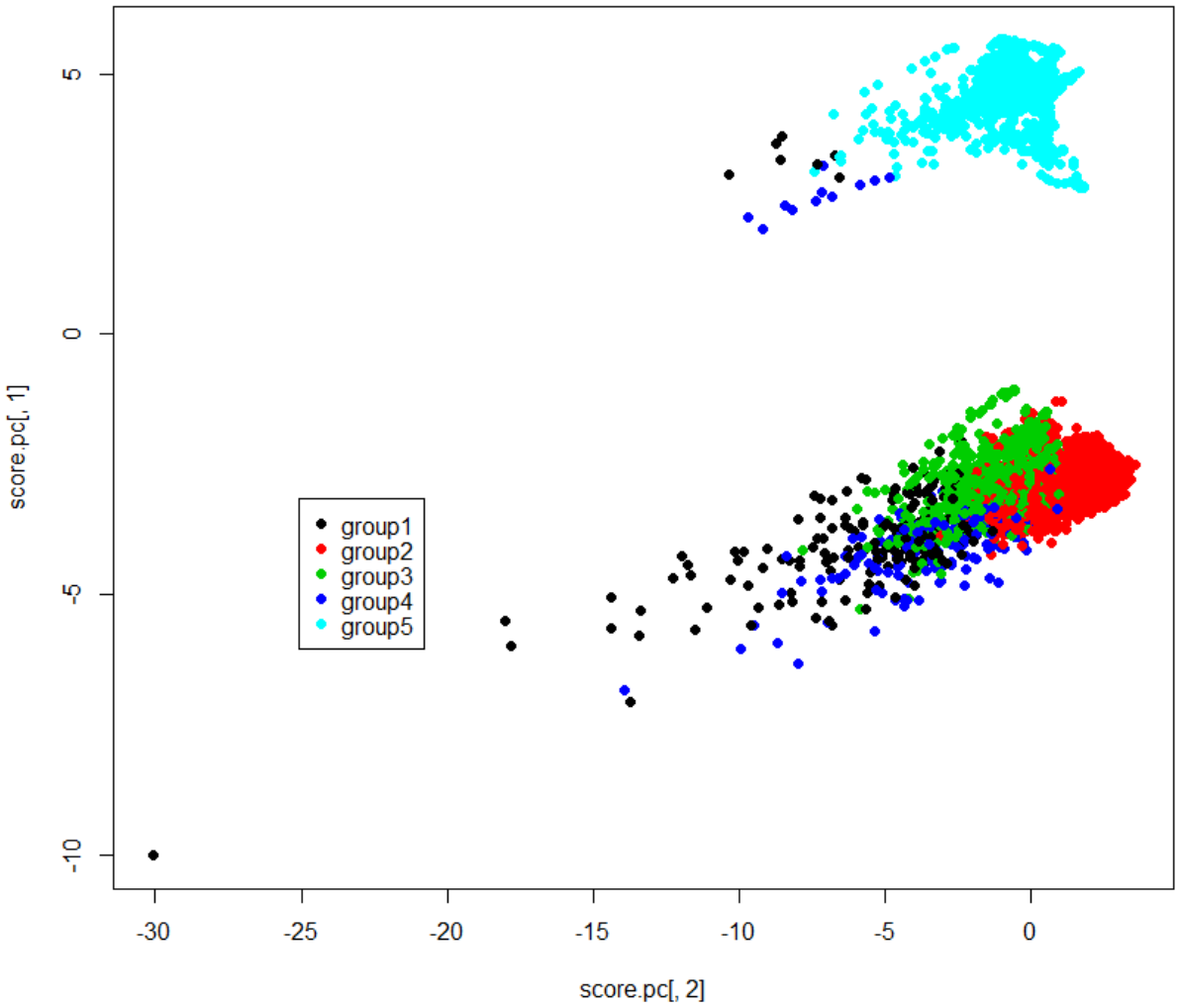
As we are required to have 5 clusters for 5 different gestures, this method might not work well because the plot shows 2 cluster is enough.

As we vasualize the result in the first two components space: (The truth is first two pc are not enough)



We can see, this 5 cluster are not well-separated. We should refine the features. I noticed that “timestamp”, which supplies with us the information when the data is extract in the video, is not useful to help us cluster without any information shows these 5 gestures happen in order, even if these 5 gestures do happen in order, with misorder in this variables in the dataset, it is still helpless. So, I decide to remove this variable.

Then we use k-means method to re-cluster, when vasualizing the result in the first two PCs space:



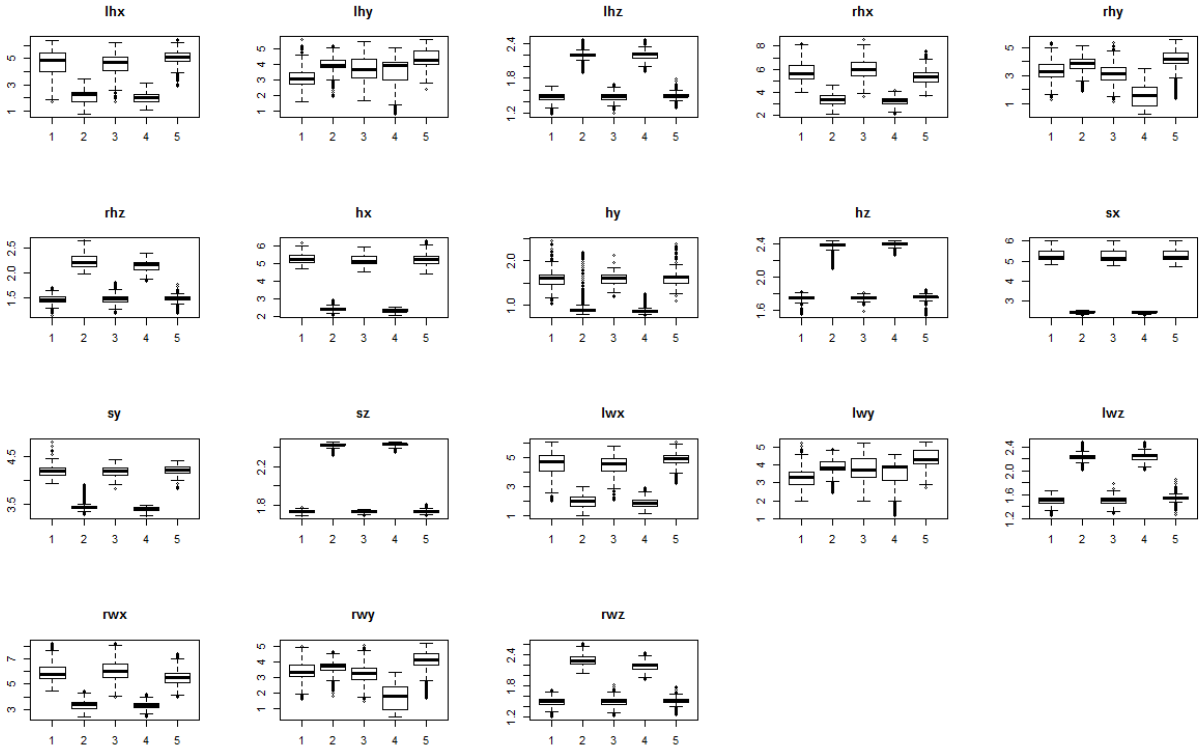
We find group 5 is very different from the other four groups, and the main difference is concentrated on the frist PC scores.

And the first PC is mainly loaded on the “position” (the first 18 variables in dataset):

```
          lhx           lhy           lhz           rhx           rhy
-0.2417084597 -0.0072874538  0.2614971768 -0.2311407169 -0.0819178854
          rhz            hx            hy            hz            sx
 0.2598748841 -0.2627265560 -0.2461443482  0.2650348062 -0.2630345582
           sy            sz           lwx           lwy           lwz
-0.2613283781  0.2654592578 -0.2478157332 -0.0355507972  0.2631782938
          rwx           rwy           rwz            x1            x2
-0.2425842904 -0.1023981336  0.2616237400 -0.0027640525  0.0009278374
           x3            x4            x5            x6            x7
-0.0034132643  0.0077841002  0.0002449098 -0.0095169833 -0.0030570035
```

```
           X8            X9           X10           X11           X12
 0.0021945549 -0.0036198671  0.0082632309  0.0010192127 -0.0087860484
          X13           X14           X15           X16           X17
-0.0022176839 -0.0034440240  0.0014804193  0.0051146377 -0.0038612380
          X18           X19           X20           X21           X22
 0.0013378632 -0.0024850820 -0.0037361397  0.0007034377  0.0060606012
          X23           X24           X25           X26           X27
-0.0031974969  0.0017398012 -0.0832430541 -0.0708609571 -0.0899092238
          X28           X29           X30           X31           X32
-0.0750980805 -0.1076225798 -0.0910250677 -0.1138118345 -0.0947405512
```

And we can draw the boxplot, which can show us the difference between group 5 and the other four groups:



We can find from the boxplot that in the range of variables like lhx, rhy are very different from other four groups. This means, postion is a very important index to separate group 5 from all 5 groups.

The units that labeled as group 5 is:

```
> group1
  [1]   2   3   4   7   8   9  10  11  12  13  15  17  18  21  23  24  26
 [18]  27  28  29  30  33  35  36  37  38  39  42  44  46  47  48  50  51
 [35]  52  55  56  57  58  59  60  61  65  67  68  70  71  72  73  74  75
```

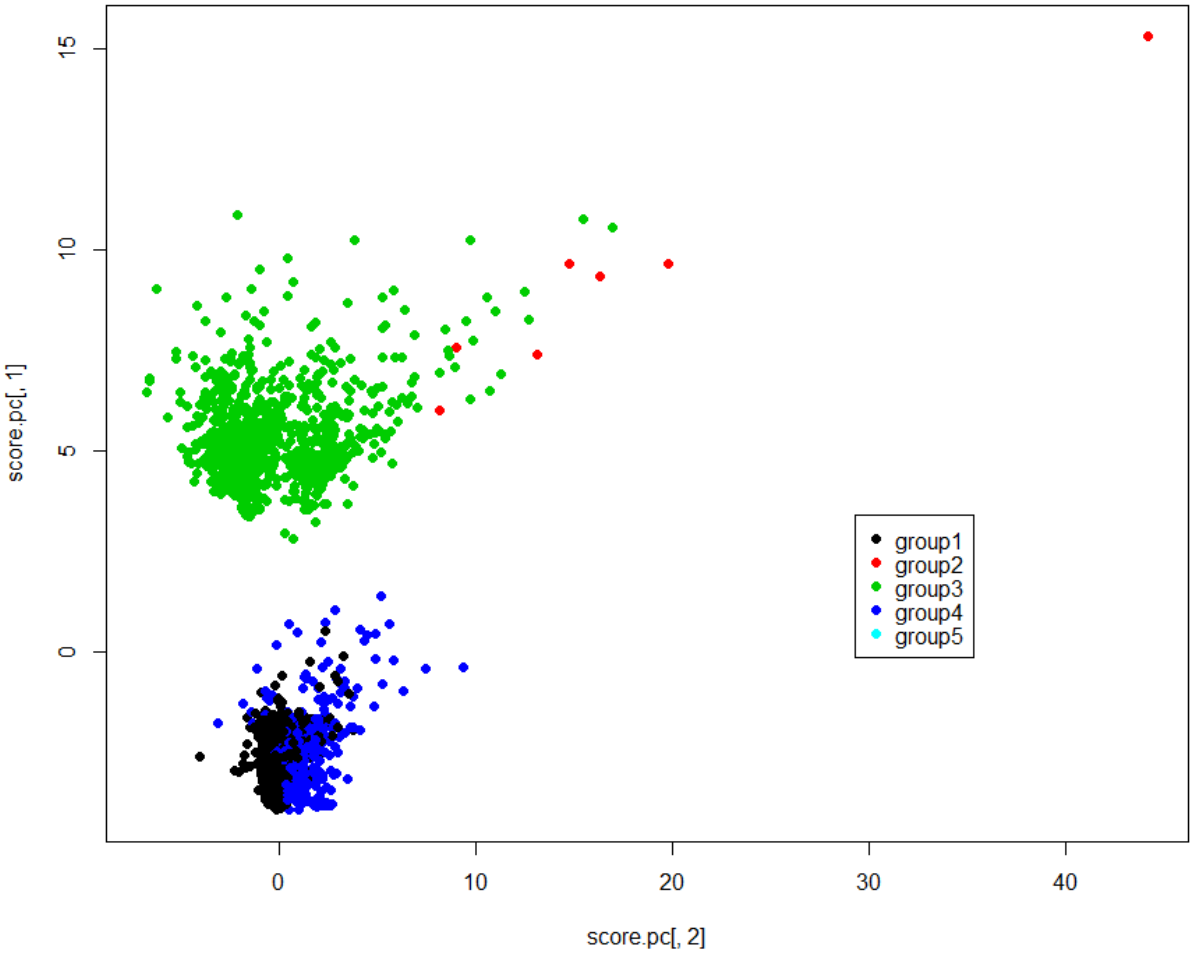
 [52]   78   81   82   83   84   85   86   87   88   90   91   93   95   96 100 101 102
 [69] 103 104 105 108 109 111 112 113 114 115 116 118 120 122 123 124 128
 [86] 129 131 132 133 134 137 138 139 141 143 144 145 146 147 148 149 150
[103] 152 154 156 158 160 161 162 163 164 165 167 169 170 171 172 173 174
[120] 175 176 177 178 180 181 186 189 191 192 193 194 195 196 198 200 201
[137] 203 204 205 206 208 209 211 212 213 216 217 218 219 220 224 225 227
[154] 229 230 232 233 234 235 236 237 238 239 240 241 242 244 245 246 247
[171] 248 251 254 255 256 257 258 259 261 262 263 265 266 267 268 271 272
[188] 275 276 277 278 279 284 285 287 288 289 290 292 293 294 295 296 298
[205] 299 303 304 305 306 307 308 309 311 313 314 317 318 319 320 322 323
[222] 324 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 342
[239] 344 345 348 352 353 356 357 358 359 360 361 362 364 365 366 368 370
[256] 371 372 373 374 375 377 378 380 382 383 384 385 389 390 391 392 395
[273] 399 401 402 403 405 407 408 409 410 413 414 415 416 417 420 421 422
[290] 423 424 425 426 428 429 431 434 435 438 439 441 442 443 444 445 446
[307] 447 448 450 451 452 453 454 456 457 458 459 460 462 464 465 469 470
[324] 474 475 476 479 480 481 483 484 487 488 490 493 494 496 497 498 503
[341] 505 506 507 508 509 510 513 514 515 516 517 518 519 520 521 522 523
[358] 524 525 527 528 531 532 533 534 535 536 537 538 539 540 541 542 543
[375] 545 546 547 548 549 550 551 553 554 555 556 557 560 563 564 565 566
[392] 568 569 570 572 573 574 575 576 577 580 582 583 585 586 587 588 590
[409] 591 594 595 596 597 598 599 600 601 602 603 604 606 607 608 609 610
[426] 611 612 613 614 615 616 617 618 620 621 622 623 624 625 626 627 628
[443] 629 630 631 633 634 635 636 637 638 640 641 643 644 645 646 647 648
[460] 649 650 651 653 654 655 656 657 658 659 660 663 664 665 667 668 669
[477] 671 676 677 678 679 680 681 682 683 684 685 686 688 689 690 691 694
[494] 695 696 697 699 702 703 704 705 706 707 708 709 711 712 714 716 717
[511] 718 719 720 721 722 723 724 725 727 731 732 735 737 738 739 740 741
[528] 742 743 745 746 747 748 752 753 755 757 761 763 764 765 767 769 770
[545] 771 772 773 775 777 778 779 780 781 782 784 785 788 789 790 792 795
[562] 796 797 798 799 802 805 806 807 808 809 810 811 812 814 816 818 820
[579] 823 825 827 828 829 830 831 832 833 834 835 836 837 838 839 840 842
[596] 843 845 846 847 848 850 851 853 854 855 857 858 859 860 863 866 867
[613] 868 869 871 872 873 874 875 876 877 878 879 880 881 882 883 886 887
[630] 888 890 891 892 893 895 896 897 899 900 901 902 904 906 907 908 909
[647] 910 912 913 914 915 916 917 918 919 920 924 927 928 929 930 931 934
[664] 936 937 938 940 942 943 944 945 947 948 949 951 952 954 955 956 957
[681] 960 961 962 963 965 966 967 969 972 973 974 976 978 979 984 985 989
[698] 993 994 995 996 997 998 999 1000 1001 1002 1003 1004 1005 1008 1009 1012 1013
[715] 1014 1015 1017 1018 1019 1022 1023 1025 1028 1029 1031 1032 1033 1034 1036 1037 1038
[732] 1039 1040 1042 1043 1044 1047 1048 1049 1050 1052 1053 1054 1056 1057 1060 1061 1062
[749] 1063 1064 1065 1069 1070 1074 1075 1077 1078 1079 1081 1084 1086 1087 1088 1089 1090
[766] 1091 1092 1093 1095 1096 1097 1099 1100 1102 1103 1104 1105 1106 1107 1109 1110 1111
[783] 1112 1113 1114 1116 1117 1118 1119 1120 1122 1124 1126 1127 1128 1129 1130 1132 1133
[800] 1134 1135 1136 1138 1139 1140 1143 1144 1147 1148 1149 1150 1153 1154 1156 1157 1158

[817] 1160 1161 1162 1163 1165 1167 1168 1170 1171 1172 1173 1174 1176 1177 1179 1181 1182
[834] 1183 1184 1185 1186 1187 1188 1189 1190 1192 1193 1194 1195 1196 1197 1198 1199 1202
[851] 1203 1204 1205 1206 1207 1209 1211 1212 1213 1214 1216 1217 1218 1219 1222 1223 1224
[868] 1226 1227 1229 1230 1231 1237 1239 1240 1241 1243 1245 1246 1248 1249 1250 1251 1252
[885] 1254 1256 1257 1258 1259 1260 1261 1262 1264 1266 1267 1268 1269 1274 1275 1276 1277
[902] 1278 1281 1284 1285 1286 1287 1288 1289 1290 1291 1292 1293 1294 1295 1296 1297 1298
[919] 1299 1301 1302 1303 1304 1305 1310 1311 1312 1313 1314 1315 1318 1320 1321 1322 1323
[936] 1324 1325 1327 1328 1331 1332 1335 1336 1341 1342 1343 1344 1345 1346 1347 1348 1349
[953] 1350 1351 1352 1353 1354 1355 1356 1357 1359 1360 1361 1362 1363 1365 1368 1369 1370
[970] 1371 1372 1373 1376 1378 1379 1381 1382 1383 1384 1386 1387 1389 1390 1391 1392 1393
[987] 1394 1398 1399 1400 1401 1402 1404 1405 1406 1407 1408 1409 1412 1416 1417 1418 1419
[1004] 1420 1421 1423 1426 1427 1428 1429 1430 1431 1432 1433 1438 1439 1440 1444 1445 1447
[1021] 1448 1449 1450 1452 1453 1454 1456 1459 1460 1462 1463 1464 1466 1470 1471 1472 1473
[1038] 1474 1475 1476 1478 1479 1480 1481 1482 1484 1485 1486 1487 1490 1494 1496 1497 1498
[1055] 1499 1500 1501 1502 1503 1504 1505 1506 1508 1509 1510 1512 1513 1514 1515 1516 1517
[1072] 1519 1521 1522 1523 1524 1525 1526 1527 1528 1529 1530 1532 1533 1536 1538 1539 1540
[1089] 1541 1545 1546 1547 1548 1549 1550 1551 1552 1554 1556 1558 1560 1562 1563 1565 1567
[1106] 1568 1569 1572 1573 1576 1577 1578 1579 1582 1583 1585 1588 1591 1592 1594 1595 1596
[1123] 1598 1599 1600 1605 1608 1610 1611 1614 1615 1616 1618 1620 1622 1623 1624 1625 1626
[1140] 1629 1631 1633 1635 1636 1639 1640 1641 1642 1643 1644 1645 1646 1647 1648 1650 1651
[1157] 1652 1653 1654 1655 1656 1657 1658 1660 1661 1663 1664 1665 1667 1668 1669 1670 1671
[1174] 1674 1675 1676 1677 1678 1679 1680 1682 1684 1685 1686 1688 1690 1691 1692 1693 1694
[1191] 1695 1697 1698 1699 1700 1702 1703 1704 1706 1707 1708 1710 1712 1713 1714 1716 1718
[1208] 1719 1720 1722 1723 1725 1726 1730 1731 1732 1733 1735 1737 1738 1739 1743 1744 1745
[1225] 1746 1747 1749 1750 1752 1755 1757 1758 1759 1760 1761 1762 1764 1765 1768 1769 1770
[1242] 1772 1773 1774 1778 1779 1780 1781 1782 1784 1785 1787 1788 1789 1790 1793 1795 1797
[1259] 1798 1799 1800 1801 1802 1803 1806 1808 1809 1810 1811 1813 1814 1815 1816 1819 1820
[1276] 1821 1822 1824 1828 1829 1830 1831 1832 1833 1834 1837 1838 1839 1840 1841 1842 1843
[1293] 1844 1845 1847 1848 1850 1853 1854 1855 1856 1857 1858 1859 1860 1864 1865 1866 1867
[1310] 1868 1869 1870 1871 1872 1874 1876 1879 1882 1885 1886 1887 1888 1890 1891 1893 1895
[1327] 1896 1897 1899 1905 1906 1907 1908 1909 1910 1912 1916 1917 1918 1919 1920 1921 1923
[1344] 1931 1932 1933 1934 1936 1937 1938 1940 1942 1946 1947 1948 1949 1950 1954 1957 1958
[1361] 1960 1961 1962 1963 1964 1965 1966 1969 1970 1972 1974 1975 1976 1978 1980 1981 1982
[1378] 1984 1985 1987 1988 1992 1993 1996 1997 1999 2000 2002 2003 2006 2007 2008 2009 2010
[1395] 2013 2016 2017 2018 2021 2024 2025 2028 2029 2032 2033 2035 2036 2038 2039 2040 2041
[1412] 2042 2044 2046 2048 2049 2050 2052 2053 2054 2055 2059 2061 2063 2064 2066 2067 2070
[1429] 2071 2072 2073 2074 2075 2076 2077 2078 2079 2081 2082 2083 2084 2085 2086 2087 2089
[1446] 2092 2094 2100 2103 2104 2105 2107 2108 2109 2110 2111 2112 2113 2114 2116 2117 2118
[1463] 2119 2120 2121 2123 2125 2128 2129 2130 2131 2135 2136 2137 2138 2139 2141 2143 2144
[1480] 2149 2151 2152 2153 2154 2155 2156 2157 2158 2159 2160 2162 2163 2164 2165 2166 2168
[1497] 2169 2170 2172 2173 2174 2175 2177 2178 2179 2180 2182 2184 2185 2186 2187 2188 2189
[1514] 2190 2192 2193 2194 2196 2197 2200 2203 2204 2206 2207 2208 2210 2213 2214 2215 2216
[1531] 2218 2220 2221 2222 2223 2224 2225 2226 2227 2229 2230 2231 2232 2235 2236 2237 2239
[1548] 2240 2241 2242 2243 2244 2246 2247 2248 2249 2250 2251 2252 2256 2257 2260 2262 2265
[1565] 2267 2268 2271 2273 2275 2276 2278 2280 2281 2282 2283 2284 2285 2286 2287 2288 2289

```
[1582] 2290 2292 2293 2294 2295 2296 2299 2300 2301 2302 2303 2304 2305 2306 2308 2312 2313
[1599] 2314 2315 2318 2319 2322 2324 2325 2327 2328 2331 2332 2333 2334 2335 2338 2339 2340
[1616] 2341 2343 2344 2345 2348 2350 2351 2353 2354 2356 2357 2362 2363 2364 2366 2367 2368
[1633] 2369 2372 2373 2374 2376 2377 2378 2380 2381 2382 2386 2388 2389 2390 2393 2396 2397
[1650] 2398 2400 2403 2404 2405 2407 2408 2409 2410 2411 2414 2415 2417 2418 2419 2420 2422
[1667] 2423 2424 2427 2428 2429 2431 2432 2433 2434 2435 2438 2439 2440 2441 2443 2446 2447
[1684] 2448 2449 2450 2452 2453 2457 2458 2459 2460 2461 2462 2463 2465 2467 2469 2470 2471
[1701] 2472 2474 2475 2477 2478 2479 2482 2483 2484 2485 2486 2487 2488 2490 2491 2492 2495
[1718] 2496 2498 2499 2502 2503 2505 2509 2511 2513 2514 2515 2516 2518 2519 2520 2522 2525
[1735] 2527 2528 2529 2530 2532 2534 2535 2537 2538 2540 2543 2544 2545 2546 2547 2549 2550
[1752] 2551 2554 2555 2557 2558 2559 2560 2561 2562 2563 2565 2566 2567 2568 2569 2570 2571
[1769] 2572 2574 2575 2577 2578 2579 2581 2582 2584 2587 2588 2589 2590 2591 2593 2595 2597
[1786] 2598 2599 2600 2601 2602 2605 2606 2609 2610 2612 2613 2614 2618 2619 2620 2621 2623
[1803] 2625 2628 2629 2632 2633 2634 2635 2636 2638 2639 2641 2642 2643 2644 2646 2651 2653
[1820] 2654 2656 2657 2659 2660 2662 2664 2668 2672 2674 2676 2678 2679 2681 2682 2683 2685
[1837] 2686 2688 2690 2691 2692 2693 2696 2699 2702 2706 2707 2708 2709 2710 2711 2713 2714
[1854] 2715 2716 2718 2722 2723 2729 2731 2732 2735 2736 2738 2739 2740 2741 2742 2743 2744
[1871] 2745 2747 2748 2750 2751 2753 2757 2763 2766 2768 2770 2773 2776 2777 2781 2783 2784
[1888] 2786 2787 2788 2789 2790 2791 2792 2793 2794 2796 2797 2799 2801 2802 2805 2806 2807
[1905] 2809 2810 2812 2813 2815 2817 2822 2823 2824 2826 2828 2831 2832 2833 2834 2836 2837
[1922] 2838 2840 2841 2842 2843 2844 2845 2846 2847 2849 2850 2851 2852 2853 2854 2855 2856
[1939] 2857 2859 2860 2861 2863 2864 2865 2866 2867 2868 2870 2871 2875 2877 2879 2880 2881
[1956] 2882 2883 2886 2887 2888 2889 2891 2893 2894 2895 2897 2898 2899 2901 2902 2904 2907
[1973] 2908 2909 2911 2912 2913 2914 2915 2919 2922 2925 2926 2929 2930 2932 2934 2935 2937
[1990] 2939 2940 2943 2945 2947 2949 2951 2952 2954 2956 2958 2960 2963 2966 2967 2969 2976
[2007] 2978 2979 2980 2981 2982 2983 2984 2985 2988 2989 2990 2991 2992 2993 2997 2998 2999
[2024] 3000 3001
```

Then we remove group 5 from the dataset to find a way to separate the other four groups.

Then I use K-means cluster and set centers equal to 4. The result shows that group 3 is very different from other 3 groups with respect to the first PC scores:



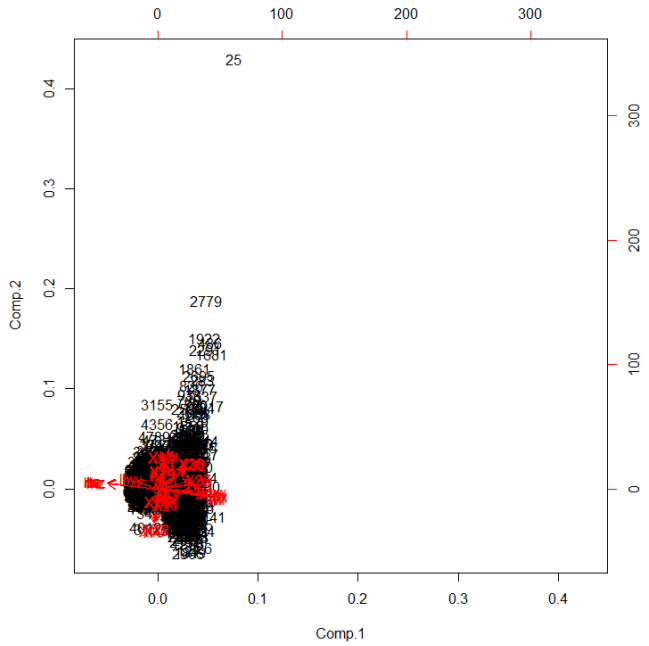
We can find group 3 is very different from other three groups in the first PC scores. And the first PC here is:

```
 lhx           lhy           lhz           rhx           rhy           rhz
0.2181025640 -0.1153474531 -0.2441302273  0.2180262158  0.0106168421 -0.2395167556
          hx            hy            hz            sx            sy            sz
0.2446772565  0.2224799627 -0.2454518269  0.2447971815  0.2423653108 -0.2470884175
         lwx           lwy           lwz           rwx           rwy           rwz
0.2256998849 -0.0972741559 -0.2461050205  0.2268392471  0.0304648069 -0.2423922841
          X1            X2            X3            X4            X5            X6
0.0368131752  0.0409056591  0.0465197761 -0.0220191958  0.0374171637  0.0250955396
          X7            X8            X9           X10           X11           X12
0.0374529005  0.0381888238  0.0512189650 -0.0190484323  0.0368764517  0.0245249578
         X13           X14           X15           X16           X17           X18
0.0178648955 -0.0006320907  0.0310561072 -0.0063927008  0.0146689847  0.0171219212
         X19           X20           X21           X22           X23           X24
```

```
0.0202384961 -0.0041407347  0.0323019415 -0.0040450230  0.0150257109  0.0165486619
         X25           X26           X27           X28           X29           X30
0.1540682742  0.1287078391  0.1587303395  0.1348550665  0.1626896522  0.1320746486
         X31           X32
0.1659443054  0.1363931714
```

(The other PCs are not shown, and the result is: the other PCs give much weight to the "vectorial velocity")

We can also see the biplot(which is not so clear):



The frist PC here gives much weight on the "position" variabels and "scalar velocity" variables. So, the "postion" and "scalar velocity" characterize the second group(which is also group 3 in this step)

The second group is:

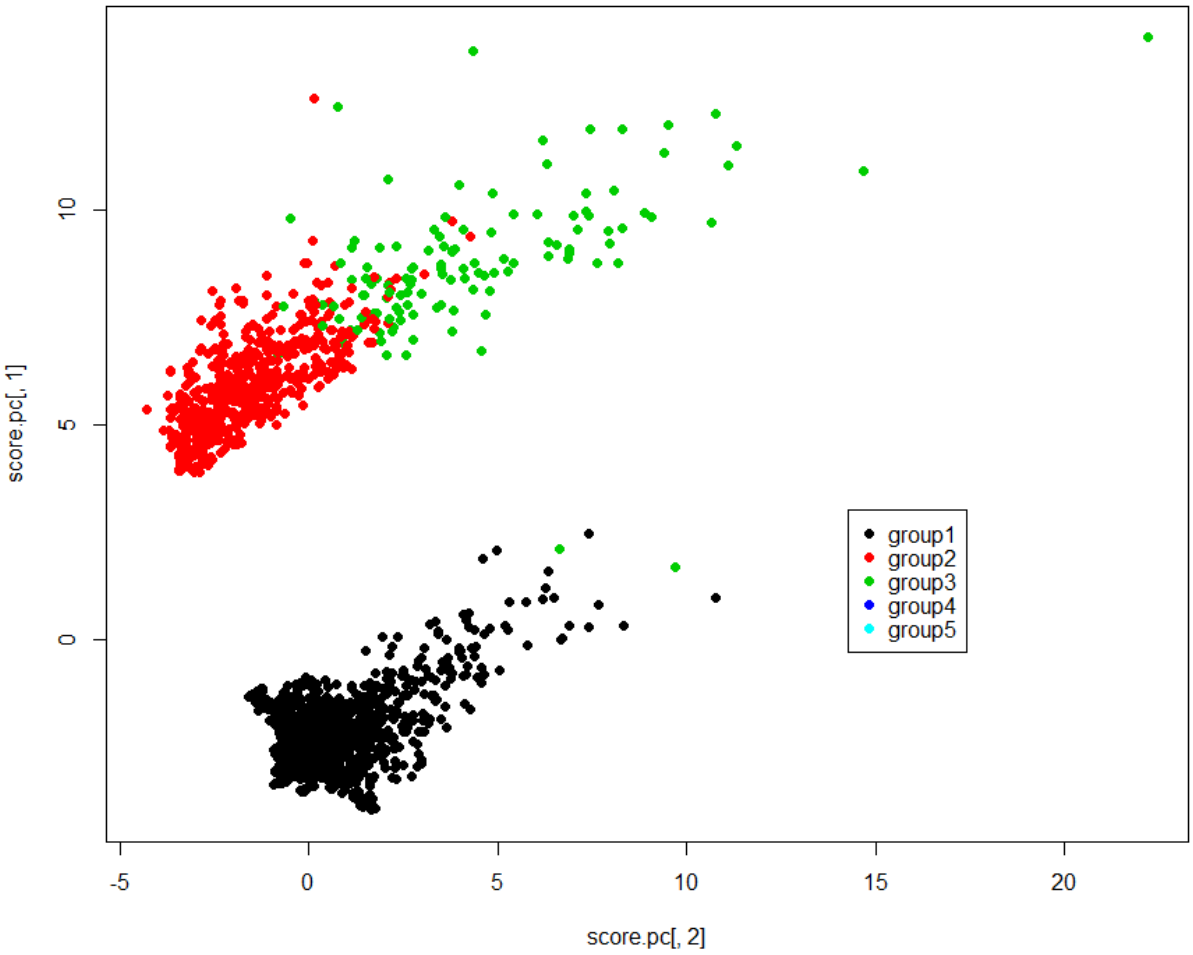
```
> group2
  [1]   14   16   19   20   25   34   43   54   64   66   69   76   98  121  125  130  135
 [18]  136  159  183  185  190  202  215  221  243  252  253  260  282  283  297  310  341
 [35]  347  349  379  381  396  400  406  418  427  432  440  466  468  478  482  486  495
 [52]  500  501  504  567  571  581  584  593  639  652  670  672  692  698  710  728  730
 [69]  750  754  758  759  760  766  768  774  776  783  787  817  826  849  862  885  911
 [86]  921  925  933  935  959  971  981  986  987  990  991 1007 1030 1035 1051 1066 1067
[103] 1071 1076 1083 1094 1142 1164 1175 1180 1220 1233 1282 1283 1309 1316 1326 1329 1333
[120] 1337 1377 1380 1395 1414 1415 1434 1442 1465 1467 1477 1489 1493 1511 1520 1537 1542
[137] 1544 1559 1566 1584 1587 1589 1593 1601 1603 1606 1612 1617 1619 1621 1627 1630 1634
[154] 1637 1662 1672 1681 1683 1689 1705 1709 1711 1717 1721 1763 1767 1771 1786 1804 1818
```

```
[171] 1826 1836 1851 1861 1877 1884 1894 1898 1904 1922 1925 1928 1939 1943 1968 1973 1977
[188] 1989 1990 1994 2004 2012 2023 2026 2027 2047 2060 2065 2088 2124 2126 2145 2147 2171
[205] 2198 2205 2234 2245 2253 2254 2259 2270 2291 2298 2329 2336 2342 2346 2349 2360 2361
[222] 2375 2379 2392 2394 2421 2426 2451 2464 2494 2507 2508 2521 2523 2524 2531 2533 2536
[239] 2542 2580 2585 2616 2617 2637 2647 2649 2650 2673 2695 2698 2704 2727 2730 2734 2755
[256] 2756 2761 2771 2775 2778 2779 2782 2808 2816 2827 2829 2830 2848 2869 2873 2896 2917
[273] 2924 2928 2936 2942 2955 2968 2970 2975 2977 2986
```

Then, we remove group 3 which generated in this step, and to separate the remaining 3 groups.

Visualiziton:

We can find that group 1 is very different from the other 2 groups with respect to the score on first PC.

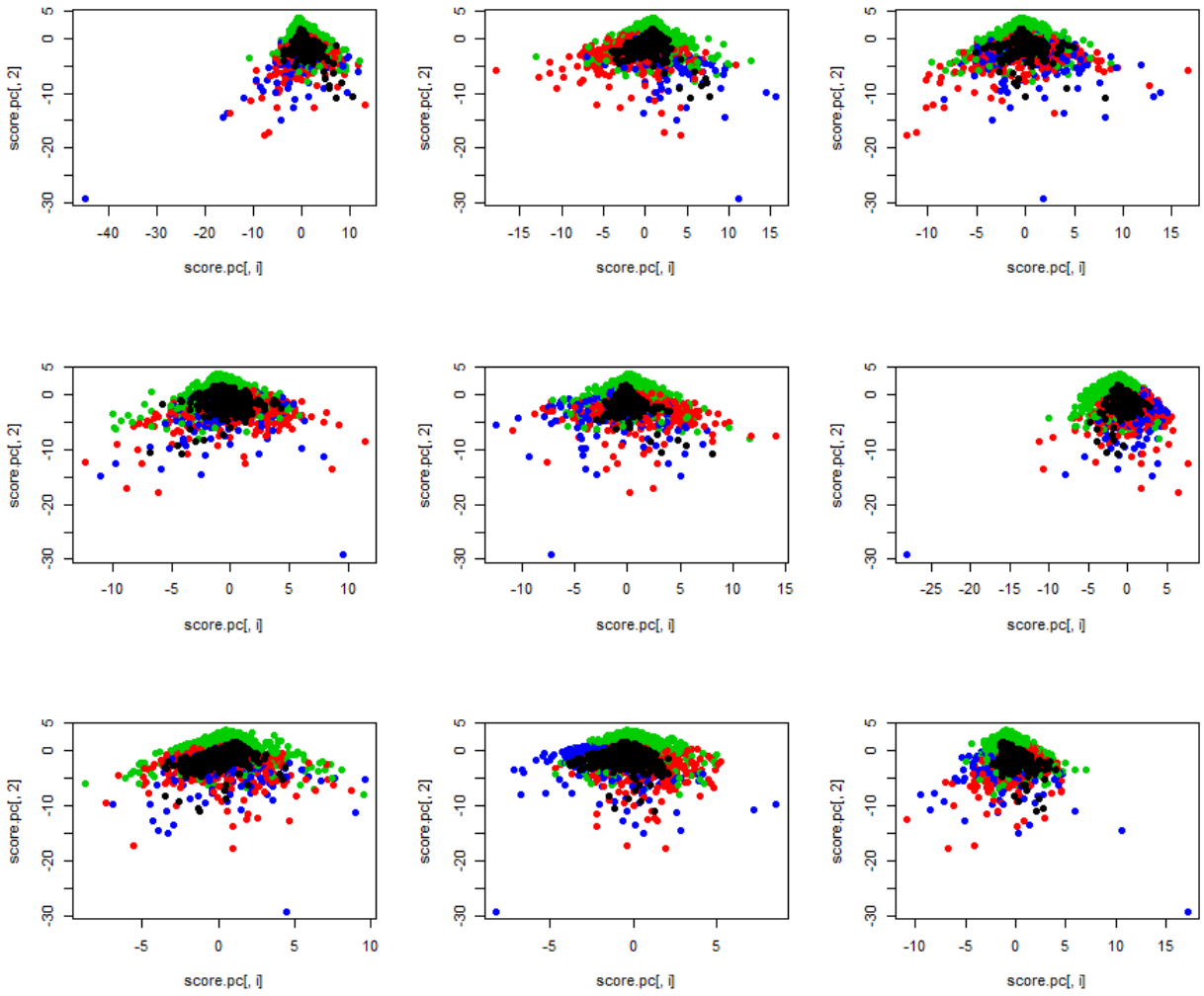


We can find group 3 is different from the other two groups with respect to second PC's score, and and second PC's is score. And when we plot the other PC's score (second PC score versus 3:11 PC score), and the PCs will shown in the appendix. All the 3:11 PCs give much weight to the "vectorial velocity" and "scalar velocity", but the second PC gives some weight to "position", some wight to "vectorial velocity", and much weight on "scalar velocity". And we can find the score on second PC is an important indicator to separate group3 from other two groups.

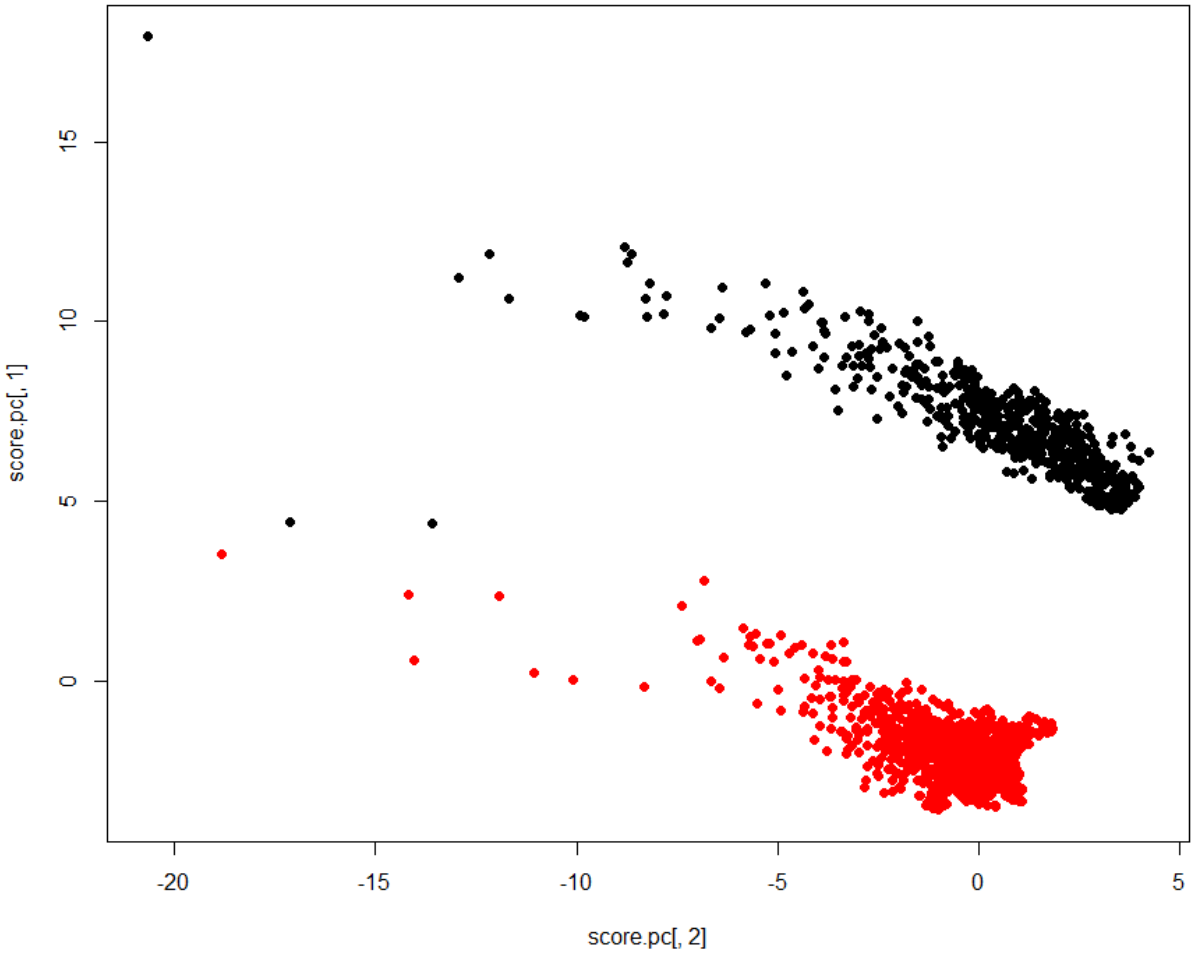
So, we can say that some "position" element, some"vectorial velocity" element, and much "scalar velocity" characterize the thrid group.

And this group is:

```
> group3
  [1]    5   53   62   89  117  127  153  168  179  197  210  226  249  264
281  302  325
 [18]  350  367  386  387  412  433  449  467  471  485  502  558  562  579
642  666  674
 [35]  687  693  700  701  713  751  786  791  821  844  852  856  898  953
968  983 1021
 [52] 1024 1027 1058 1068 1108 1121 1131 1137 1146 1152 1228 1244 1253 1270 1
413 1425 1446
 [69] 1535 1543 1555 1580 1602 1604 1701 1741 1796 1849 1852 1878 1880 1883 1
902 1952 1959
 [86] 1986 1995 2011 2015 2030 2034 2058 2097 2099 2132 2140 2150 2233 2264 2
279 2297 2337
[103] 2391 2395 2573 2594 2608 2611 2640 2645 2665 2694 2724 2726 2733 2754 2
885 2906 2918
[120] 2920 2933 2964 2974 2987 3242 3370
```



Then we contimue out process, to separate the last two group: visualize it:



We can see the first and the second PC scores both contribute to the difference between this two group.

And first two PCs in this step do not have a very meaningful description. Maybe this two group have much in common, but the main difference is the “scalar velocity”.

The fourth group is:

```
> group4
  [1]   1   6  22  31  32  40  41  45  49  63  77  79  80  92
 94  97  99
 [18] 106 107 110 119 126 140 142 151 155 157 166 182 184 187
188 199 207
 [35] 214 222 223 228 231 250 269 270 273 274 280 286 291 300
301 312 315
```

 [52] 316 321 343 346 351 354 355 363 369 376 388 393 394 397
398 404 411
 [69] 419 430 436 437 455 461 463 472 473 477 489 491 492 499
511 512 526
 [86] 529 530 544 552 559 561 578 589 592 605 619 632 661 662
673 675 715
[103] 726 729 733 734 736 744 749 756 762 793 794 800 801 803
804 813 815
[120] 819 822 824 841 861 864 865 870 884 889 894 903 905 922
923 926 932
[137] 939 941 946 950 958 964 970 975 977 980 982 988 992 1006 1
010 1011 1016
[154] 1020 1026 1041 1045 1046 1055 1059 1072 1073 1080 1082 1085 1098 1101 1
115 1123 1125
[171] 1141 1145 1151 1155 1159 1166 1169 1178 1191 1200 1201 1208 1210 1215 1
221 1225 1232
[188] 1234 1235 1236 1238 1242 1247 1255 1263 1265 1271 1272 1273 1279 1280 1
300 1306 1307
[205] 1308 1317 1319 1330 1334 1338 1339 1340 1358 1364 1366 1367 1374 1375 1
385 1388 1396
[222] 1397 1403 1410 1411 1422 1424 1435 1436 1437 1441 1443 1451 1455 1457 1
458 1461 1468
[239] 1469 1483 1488 1491 1492 1495 1507 1518 1531 1534 1553 1557 1561 1564 1
570 1571 1574
[256] 1575 1581 1586 1590 1597 1607 1609 1613 1628 1632 1638 1649 1659 1666 1
673 1687 1696
[273] 1715 1724 1727 1728 1729 1734 1736 1740 1742 1748 1751 1753 1754 1756 1
766 1775 1776
[290] 1777 1783 1791 1792 1794 1805 1807 1812 1817 1823 1825 1827 1835 1846 1
862 1863 1873
[307] 1875 1881 1889 1892 1900 1901 1903 1911 1913 1914 1915 1924 1926 1927 1
929 1930 1935
[324] 1941 1944 1945 1951 1953 1955 1956 1967 1971 1979 1983 1991 1998 2001 2
005 2014 2019
[341] 2020 2022 2031 2037 2043 2045 2051 2056 2057 2062 2068 2069 2080 2090 2
091 2093 2095
[358] 2096 2098 2101 2102 2106 2115 2122 2127 2133 2134 2142 2146 2148 2161 2
167 2176 2181
[375] 2183 2191 2195 2199 2201 2202 2209 2211 2212 2217 2219 2228 2238 2255 2
258 2261 2263
[392] 2266 2269 2272 2274 2277 2307 2309 2310 2311 2316 2317 2320 2321 2323 2
326 2330 2347
[409] 2352 2355 2358 2359 2365 2370 2371 2383 2384 2385 2387 2399 2401 2402 2
406 2412 2413

```
[426] 2416 2425 2430 2436 2437 2442 2444 2445 2454 2455 2456 2466 2468 2473 2
476 2480 2481
[443] 2489 2493 2497 2500 2501 2504 2506 2510 2512 2517 2526 2539 2541 2548 2
552 2553 2556
[460] 2564 2576 2583 2586 2592 2596 2603 2604 2607 2615 2622 2624 2626 2627 2
630 2631 2648
[477] 2652 2655 2658 2661 2663 2666 2667 2669 2670 2671 2675 2677 2680 2684 2
687 2689 2697
[494] 2700 2701 2703 2705 2712 2717 2719 2720 2721 2725 2728 2737 2746 2749 2
752 2758 2759
[511] 2760 2762 2764 2765 2767 2769 2772 2774 2780 2785 2795 2798 2800 2803 2
804 2811 2814
[528] 2818 2819 2820 2821 2825 2835 2839 2858 2862 2872 2874 2876 2878 2884 2
890 2892 2900
[545] 2903 2905 2910 2916 2921 2923 2927 2931 2938 2941 2944 2946 2948 2950 2
953 2957 2959
[562] 2961 2962 2965 2971 2972 2973 2994 2995 2996 3002 3003 3572 4158
```

And the last group is the remaining part.