Problem #1

BASIC EXPLORATORY DATA ANALYSIS

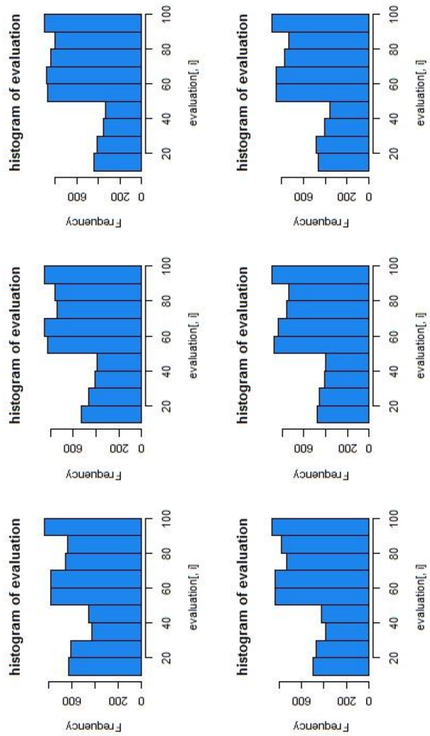
When I get the data set, I need to do some basic data explorations, such as checking the dimensionality of the data set, checking the range of each variable, checking the pairwise correlation between each pair of variables, and so on. Therefore, I first check the dimensionality of the data set, which is a 5820*33 matrix. Then, I obtain the summary of the data set, which is shown in Figure 1 below,

Figure 1 summary of the data set

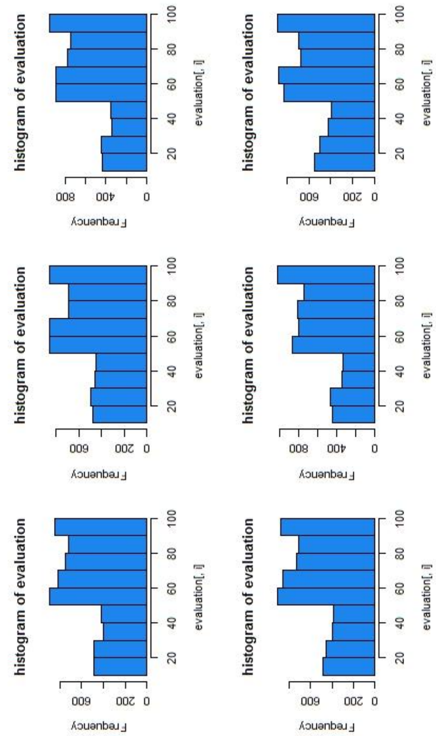
```
> summary(score)
     instr           class          nb.repeat       attendance      difficulty          Q1
 Min.   :1.000   Min.   : 1.000   Min.   :1.000   Min.   :0.000   Min.   : 15.00   Min.   :15.00
 1st Qu.:2.000   1st Qu.: 4.000   1st Qu.:1.000   1st Qu.:0.000   1st Qu.: 23.90   1st Qu.:37.41
 Median :3.000   Median : 7.000   Median :1.000   Median :1.000   Median : 59.40   Median :60.00
 Mean   :2.486   Mean   : 7.276   Mean   :1.214   Mean   :1.676   Mean   : 55.43   Mean   :58.24
 3rd Qu.:3.000   3rd Qu.:10.000   3rd Qu.:1.000   3rd Qu.:3.000   3rd Qu.: 78.48   3rd Qu.:80.08
 Max.   :3.000   Max.   :13.000   Max.   :3.000   Max.   :4.000   Max.   :100.00   Max.   :99.99
       Q2               Q3               Q4              Q5              Q6              Q7
 Min.   : 15.01   Min.   : 15.01   Min.   :15.00   Min.   :15.00   Min.   :15.01   Min.   :15.00
 1st Qu.: 40.88   1st Qu.: 43.66   1st Qu.:41.32   1st Qu.:41.49   1st Qu.:41.32   1st Qu.:40.97
 Median : 61.54   Median : 62.73   Median :61.87   Median :61.92   Median :61.97   Median :61.41
 Mean   : 61.12   Mean   : 63.12   Mean   :61.33   Mean   :61.71   Mean   :61.74   Mean   :60.91
 3rd Qu.: 81.04   3rd Qu.: 81.42   3rd Qu.:81.13   3rd Qu.:81.28   3rd Qu.:81.09   3rd Qu.:80.90
 Max.   :100.00   Max.   :100.00   Max.   :99.99   Max.   :99.99   Max.   :99.99   Max.   :99.99
       Q8               Q9              Q10              Q11             Q12             Q13
 Min.   : 15.02   Min.   :15.00   Min.   : 15.00   Min.   :15.01   Min.   :15.01   Min.   : 15.04
 1st Qu.: 40.32   1st Qu.:43.36   1st Qu.: 41.51   1st Qu.:43.04   1st Qu.:39.76   1st Qu.: 44.95
 Median : 60.96   Median :62.54   Median : 61.51   Median :62.84   Median :61.29   Median : 63.79
 Mean   : 60.47   Mean   :62.94   Mean   : 61.36   Mean   :63.19   Mean   :60.33   Mean   : 64.41
 3rd Qu.: 80.66   3rd Qu.:81.52   3rd Qu.: 80.94   3rd Qu.:82.10   3rd Qu.:80.76   3rd Qu.: 82.20
 Max.   :100.00   Max.   :99.99   Max.   :100.00   Max.   :99.99   Max.   :99.99   Max.   :100.00
      Q14              Q15              Q16             Q17              Q18             Q19
 Min.   : 15.03   Min.   : 15.02   Min.   :15.00   Min.   : 15.08   Min.   :15.01   Min.   :15.02
 1st Qu.: 55.54   1st Qu.: 55.58   1st Qu.:42.76   1st Qu.: 56.39   1st Qu.:44.28   1st Qu.:55.21
 Median : 64.52   Median : 64.25   Median :62.64   Median : 75.92   Median :63.36   Median :63.88
 Mean   : 65.39   Mean   : 65.38   Mean   :62.93   Mean   : 67.44   Mean   :63.98   Mean   :64.84
 3rd Qu.: 82.64   3rd Qu.: 82.65   3rd Qu.:82.00   3rd Qu.: 83.81   3rd Qu.:82.32   3rd Qu.:82.60
 Max.   :100.00   Max.   :100.00   Max.   :99.99   Max.   :100.00   Max.   :99.99   Max.   :99.99

      Q20             Q21              Q22              Q23             Q24              Q25

 Min.   :15.00   Min.   : 15.01   Min.   : 15.01   Min.   :15.03   Min.   : 15.02   Min.   : 15.01

 1st Qu.:55.24   1st Qu.: 55.53   1st Qu.: 55.63   1st Qu.:44.12   1st Qu.: 42.88   1st Qu.: 55.60

 Median :64.18   Median : 64.43   Median : 64.62   Median :62.99   Median : 62.50   Median : 64.49

 Mean   :65.16   Mean   : 65.68   Mean   : 65.89   Mean   :63.58   Mean   : 62.96   Mean   : 65.77

 3rd Qu.:82.97   3rd Qu.: 83.04   3rd Qu.: 83.03   3rd Qu.:82.19   3rd Qu.: 81.89   3rd Qu.: 82.94

 Max.   :99.99   Max.   :100.00   Max.   :100.00   Max.   :99.99   Max.   :100.00   Max.   :100.00

      Q26              Q27             Q28
 Min.   : 15.01   Min.   :15.03   Min.   :15.00
 1st Qu.: 44.44   1st Qu.:42.19   1st Qu.:55.56
 Median : 63.32   Median :62.71   Median :64.57
 Mean   : 64.03   Mean   :62.73   Mean   :65.69
 3rd Qu.: 82.37   3rd Qu.:81.96   3rd Qu.:83.10
 Max.   :100.00   Max.   :99.99   Max.   :99.99
```

Observing the summary, I notice that the data for evaluations are included in column 6 to column 33 in the data set. The first five columns include the information of instructor's ID, course code, number of times taking the course, level of attendance, level of difficulty, which are irrelevant to our data analysis, so I omit the first five columns and compress the data set into a 5820*28 matrix. Then, I use the histograms and box plots to check the raw distributions of each variable, as shown in Figure 2 and Figure 3 below,

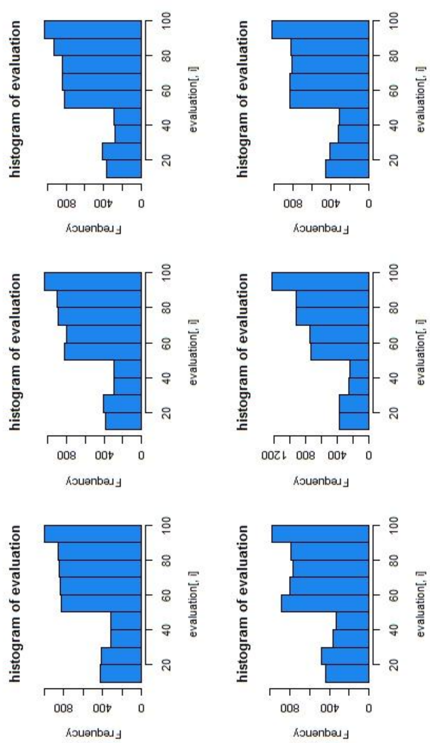
Figure 2 histograms



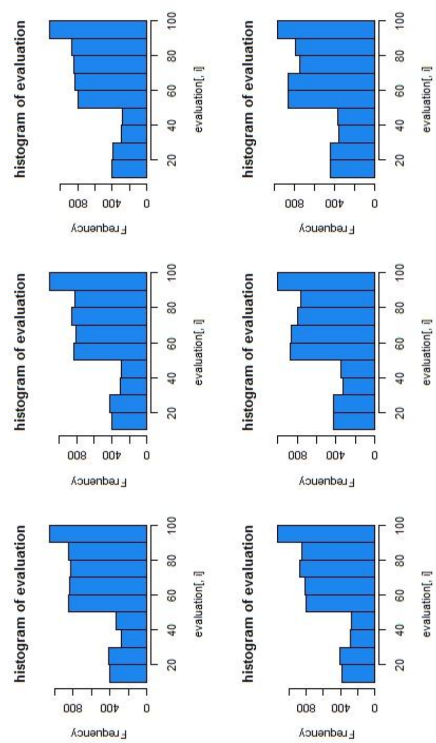
Q1 – Q6



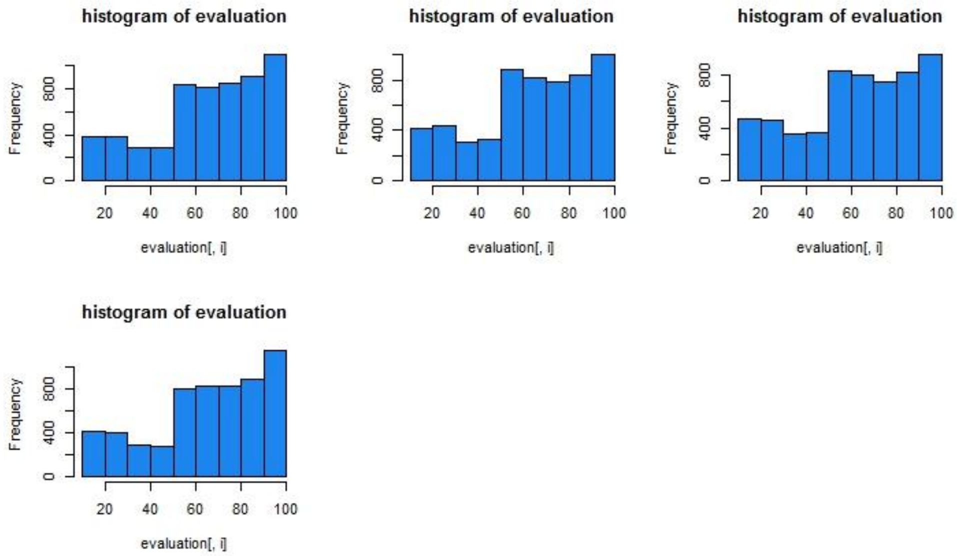
Q7 – Q12



Q13 – Q18

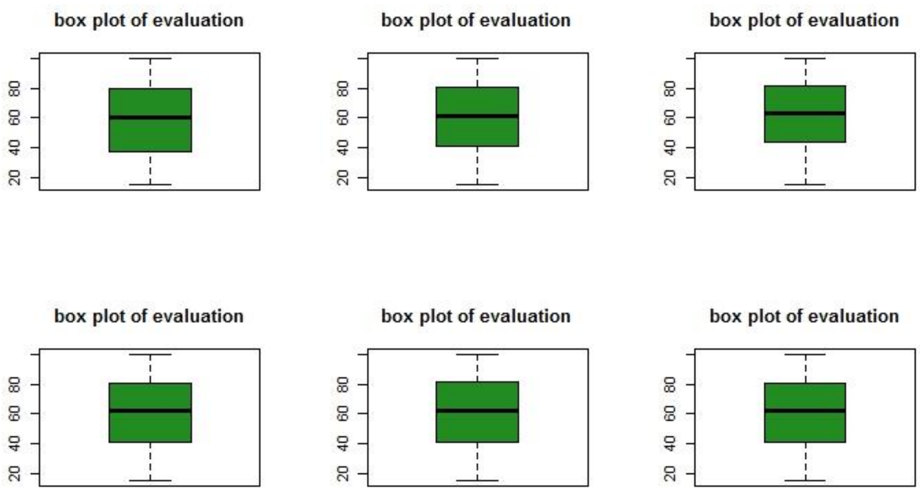


Q19 – Q24

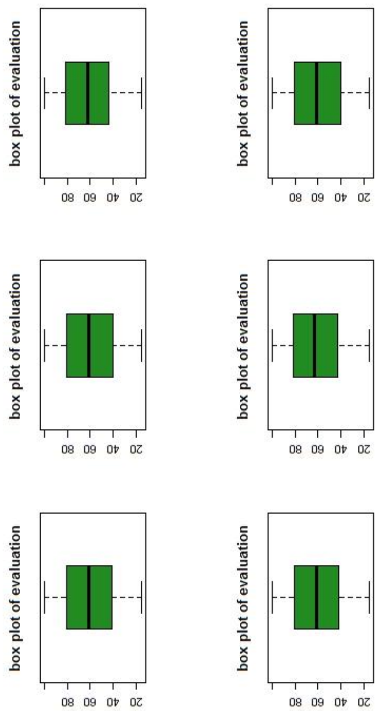


Q25 – Q28

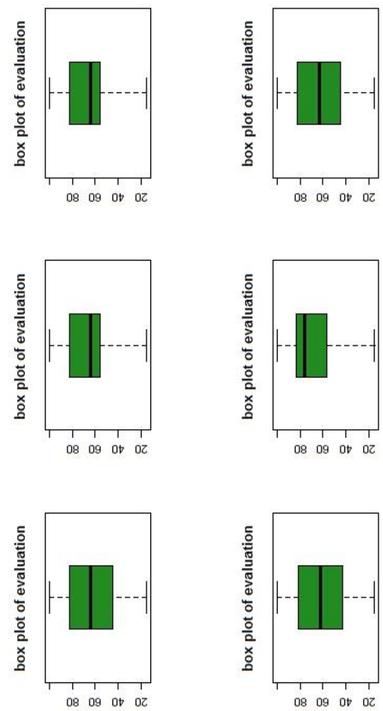
Figure 3 box plots



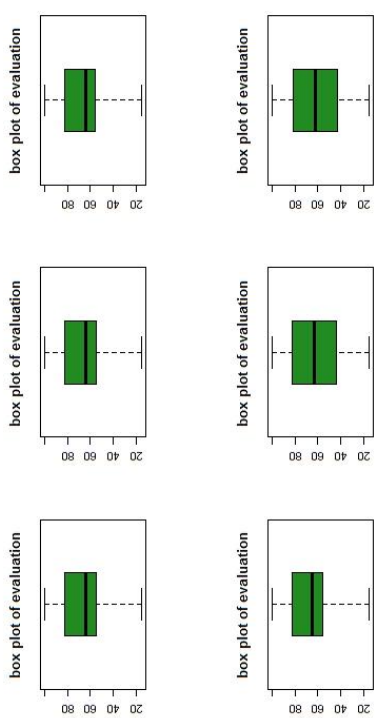
Q1 – Q6



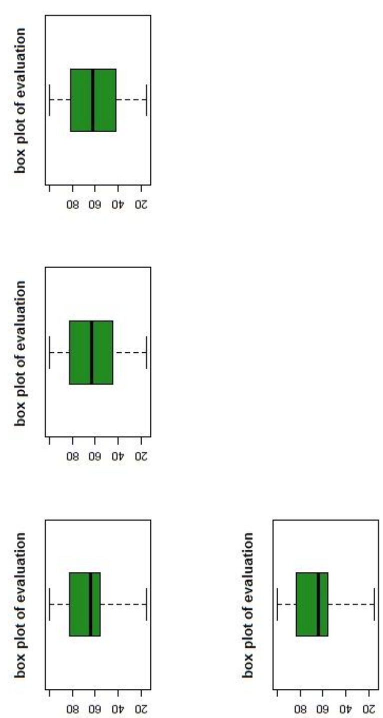
Q7 – Q12



Q13 – Q18



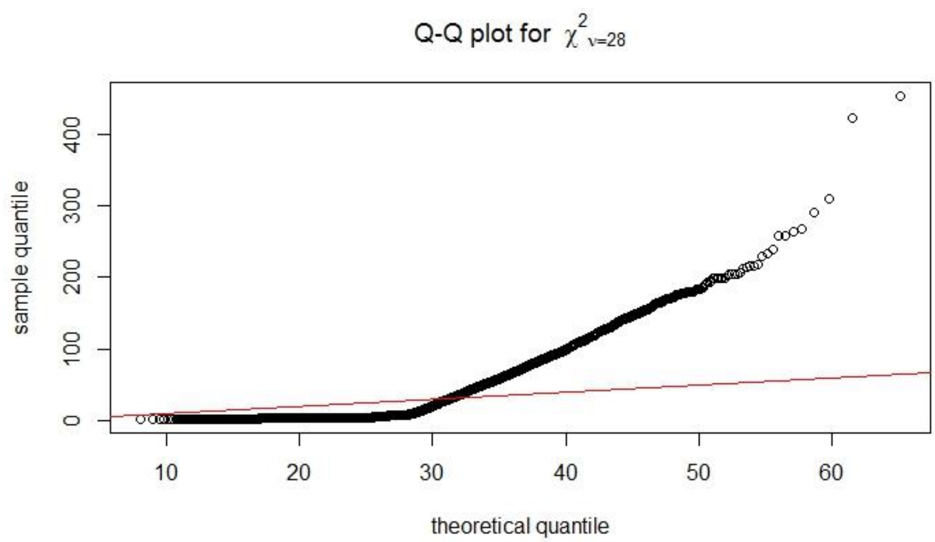
Q19 – Q24



Q25 – Q28

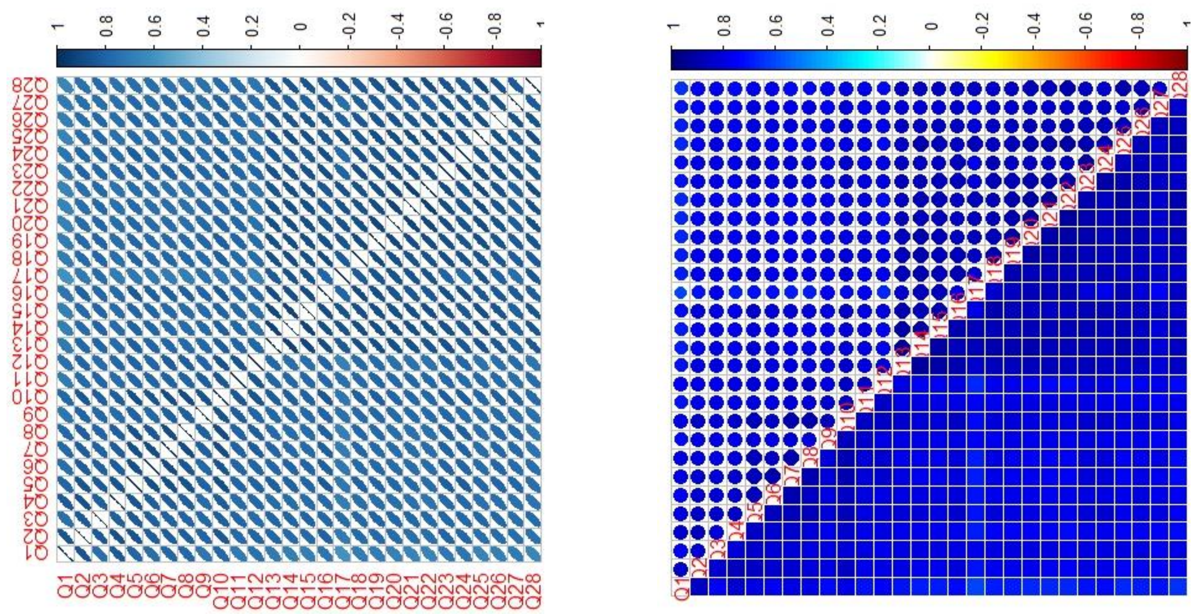
From the histograms and box pots, I find that the evaluation patterns are similar for all evaluations: the medians are within 60 to 65, the scores above 50 and the scores below 50 are approximately uniformly distributed. Next, I check the multivariate normality and outliers using the QQ plot, as shown in Figure 4 below,

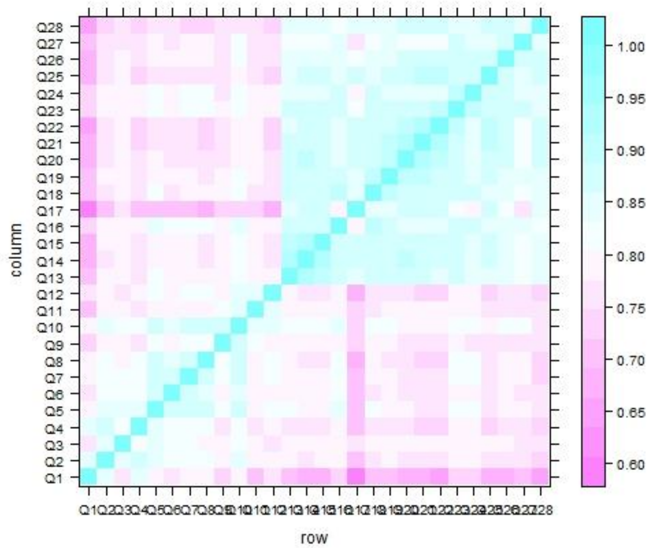
Figure 4 QQ plot



From the QQ plot, I find that the data are not normally distributed and there are few outliers in the data set. Then, I am ready for checking the dependency pattern. For checking the dependency pattern, I first compute the correlation matrix of the data set to get pairwise correlations and use corrplot() and levelplot() functions to visualize the correlation patterns, which are shown in Figure 5 below,

Figure 5 correlation matrices





According to the correlation plots above, I observe that all evaluations have relatively strong positive linear correlations, and among them, evaluations Q13 to Q28 have even stronger positive linear correlations (Q13 to Q18 are evaluations related to instructors).

<u>PRINCIPAL COMPONENT ANALYSIS (PCA)</u>

Then, I apply Principal Component Analysis to the data set to find independent orthogonal principal components. By applying PCA, I am trying to reduce the dimension and at the same time identify how different variables work together. The results are shown below in Figure 6, Figure 7 and Figure 8,

Figure 6 proportion of variance

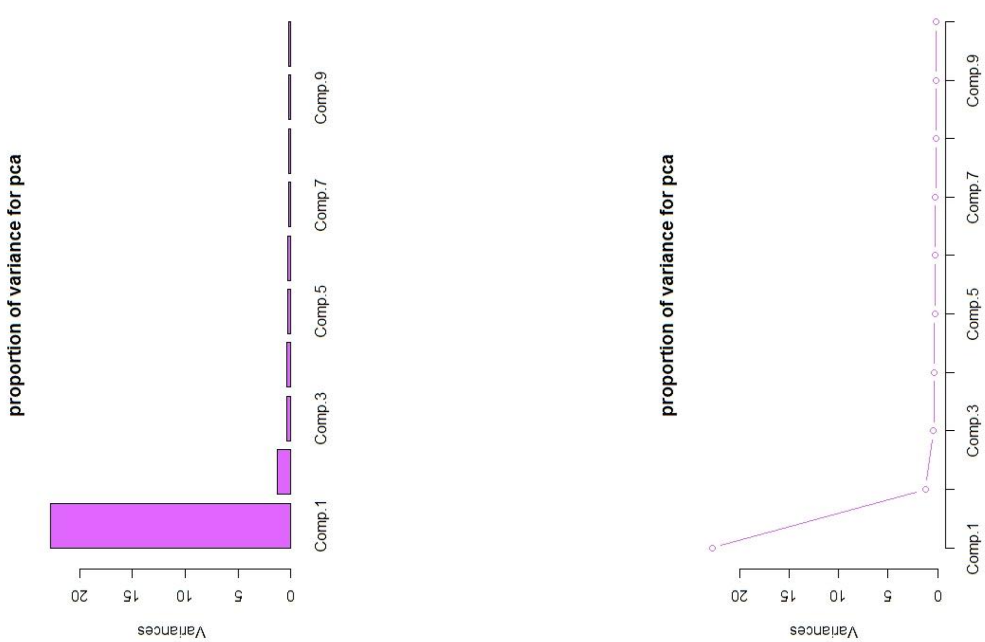


Figure 7 biplot

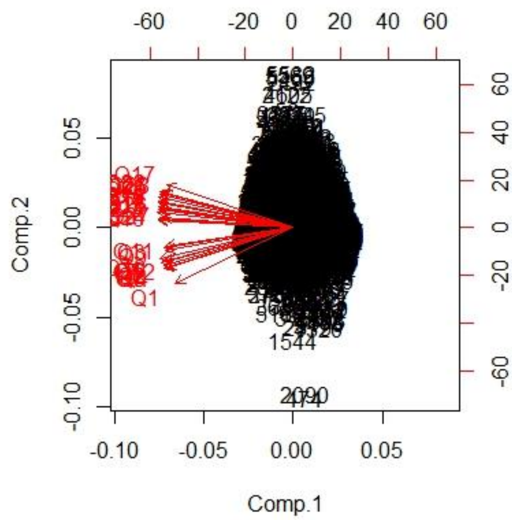


Figure 8(a) PCA results

```
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4    Comp.5      Comp.6      Comp.7
Standard deviation     4.771683 1.11810184 0.63752690 0.60806923 0.5467302 0.516616916 0.463436412
Proportion of Variance 0.813177 0.04464828 0.01451573 0.01320529 0.0106755 0.009531894 0.007670475
Cumulative Proportion  0.813177 0.85782525 0.87234098 0.88554627 0.8962218 0.905753667 0.913424142
                            Comp.8      Comp.9     Comp.10     Comp.11     Comp.12     Comp.13
Standard deviation     0.437079420 0.428820032 0.391930590 0.386700081 0.385251718 0.362439769
Proportion of Variance 0.006822801 0.006567379 0.005486057 0.005340605 0.005300675 0.004691521
Cumulative Proportion  0.920246943 0.926814322 0.932300379 0.937640985 0.942941659 0.947633180
                           Comp.14     Comp.15     Comp.16     Comp.17     Comp.18     Comp.19
Standard deviation     0.357195455 0.351909841 0.345581073 0.340760720 0.333228282 0.325131446
Proportion of Variance 0.004556735 0.004422876 0.004265224 0.004147067 0.003965753 0.003775373
Cumulative Proportion  0.952189916 0.956612792 0.960878016 0.965025083 0.968990836 0.972766209
                           Comp.20     Comp.21     Comp.22   Comp.23     Comp.24     Comp.25
Standard deviation     0.323240297 0.309976255 0.305809377 0.3015739 0.297139201 0.284954444
Proportion of Variance 0.003731582 0.003431617 0.003339978 0.0032481 0.003153275 0.002899966
Cumulative Proportion  0.976497791 0.979929408 0.983269386 0.9865175 0.989670761 0.992570727
                           Comp.26     Comp.27     Comp.28
Standard deviation     0.279497311 0.260040201 0.249559594
Proportion of Variance 0.002789955 0.002415032 0.002224285
Cumulative Proportion  0.995360682 0.997775715 1.000000000
```

From Figure 6, I am suggested to choose 2 principal components since the “elbow” is at the component 2, and from Figure 8(a), I can get the same suggestion since the cumulative proportion is greater than 85% for two principal components. Next, I check the biplot: on the plot of first two principal components, all variables are correlated with each other since the angles are small and all variables have similar variances since the lengths are similar. After checking the loadings, I observe that variables actually do not have strong correlations with the first two principal components, which means the first two principal components do not sufficiently explain all variables.

Figure 8(b) PCA results

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13
Q1	-0.170	-0.339	-0.469		-0.156	0.392	0.102		-0.237	-0.250		0.305	
Q2	-0.185	-0.235	-0.326	0.127		0.128			0.102		0.239		-0.371
Q3	-0.185	-0.121	-0.155	0.343	-0.127	-0.251	-0.174	0.285	0.668			0.142	0.119
Q4	-0.183	-0.248	-0.354				-0.107	-0.122	-0.169	0.220	-0.377	-0.584	0.192
Q5	-0.190	-0.214			0.196	-0.229			0.113	0.151		-0.203	0.222
Q6	-0.186	-0.204			0.227	-0.438	-0.209	-0.127	-0.231		0.262		-0.453
Q7	-0.187	-0.250	0.114	-0.114	0.162	-0.291		-0.161	-0.110	-0.213		0.136	
Q8	-0.186	-0.250	0.172	-0.166		-0.162	0.223			-0.397	-0.217	0.191	0.307
Q9	-0.183	-0.137	0.309	0.255	-0.150		0.668	0.245	-0.122	0.299			
Q10	-0.192	-0.194	0.217				0.150				0.159		
Q11	-0.184	-0.113	0.420	0.294	-0.252	0.152	-0.312						
Q12	-0.182	-0.211	0.372		-0.143	0.394	-0.383	-0.120			-0.124		-0.145
Q13	-0.194	0.109			0.326	0.166	-0.162	0.175	-0.132	0.140		0.243	0.312
Q14	-0.195	0.161		0.138	0.287		-0.102	0.157	-0.135	0.101		0.156	0.229
Q15	-0.194	0.158		0.130	0.272			0.106	-0.153	0.106	-0.148		
Q16	-0.195			-0.179	0.328	0.208		0.165		0.137			-0.204
Q17	-0.183	0.264		0.381					-0.103	-0.485	-0.348		-0.264
Q18	-0.193	0.127			0.280	0.220	0.129		0.271	-0.211		-0.261	-0.118
Q19	-0.194	0.152				0.119		-0.158	0.172	-0.270	0.224	-0.292	
Q20	-0.194	0.192						-0.290			0.257		0.215
Q21	-0.193	0.219			-0.161			-0.327		0.108	0.245		0.196
Q22	-0.192	0.223			-0.171			-0.306		0.131	0.108	0.114	
Q23	-0.196			-0.271				-0.211	0.180	0.196	-0.157	0.135	-0.117
Q24	-0.193			-0.369				-0.122	0.238	0.176	-0.367	0.162	-0.159
Q25	-0.192	0.208			-0.169	-0.122					-0.282	0.231	
Q26	-0.192	0.120		-0.225	-0.189	-0.150	-0.102	0.314					
Q27	-0.188			-0.417	-0.282			0.427	-0.101	-0.166	0.197	-0.242	0.108
Q28	-0.189	0.210			-0.237	-0.184	-0.132	0.118	-0.275				

FACTOR ANALYSIS (FA)

Since PCA do not give me satisfactory results, I then apply Factor Analysis to the data set to find factors or latent variables from all variables. The results are shown below in Figure 9 and Figure 10,

Figure 9(a) factor loadings (varimax)

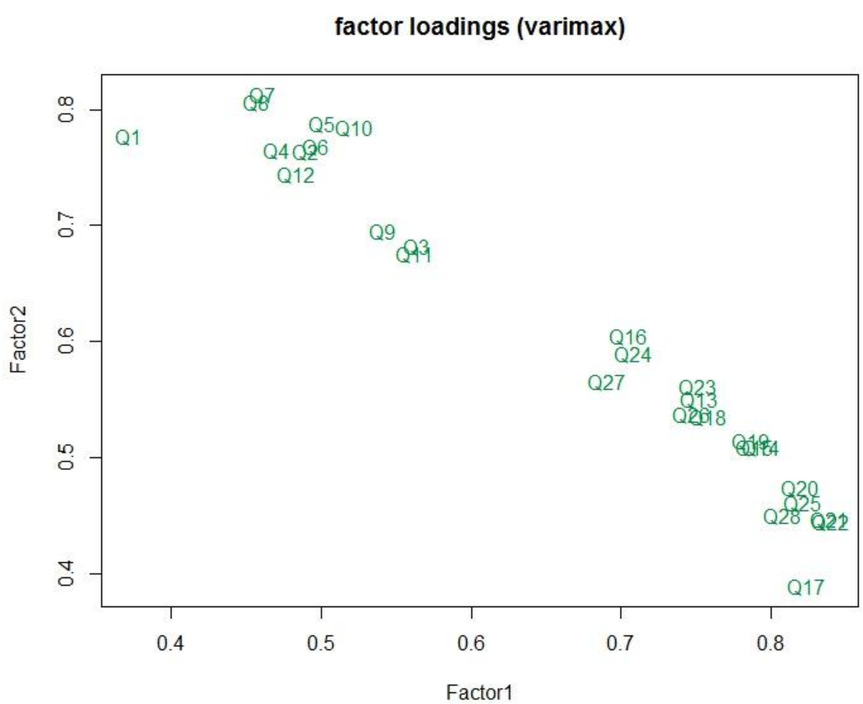


Figure 9(b) FA results (varimax)

```
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
```

```
Loadings:
    Factor1 Factor2 Factor3
Q1   0.357   0.784
Q2   0.473   0.773
Q3   0.549   0.693
Q4   0.455   0.774
Q5   0.484   0.798
Q6   0.479   0.777
Q7   0.444   0.822
Q8   0.441   0.817
Q9   0.525   0.706
Q10  0.506   0.795
Q11  0.547   0.687
Q12  0.467   0.755
Q13  0.739   0.560   0.207
Q14  0.783   0.517   0.227
Q15  0.777   0.520   0.182
Q16  0.690   0.618
Q17  0.813   0.403   0.113
Q18  0.744   0.550
Q19  0.773   0.531
Q20  0.808   0.492
Q21  0.830   0.464
Q22  0.833   0.462
Q23  0.743   0.578
Q24  0.700   0.606
Q25  0.814   0.478
Q26  0.738   0.553
Q27  0.683   0.582  -0.116
Q28  0.799   0.467

               Factor1 Factor2 Factor3
SS loadings     12.206  11.488   0.215
Proportion Var   0.436   0.410   0.008
Cumulative Var   0.436   0.846   0.854
```

Figure 10(a) factor loadings (promax)

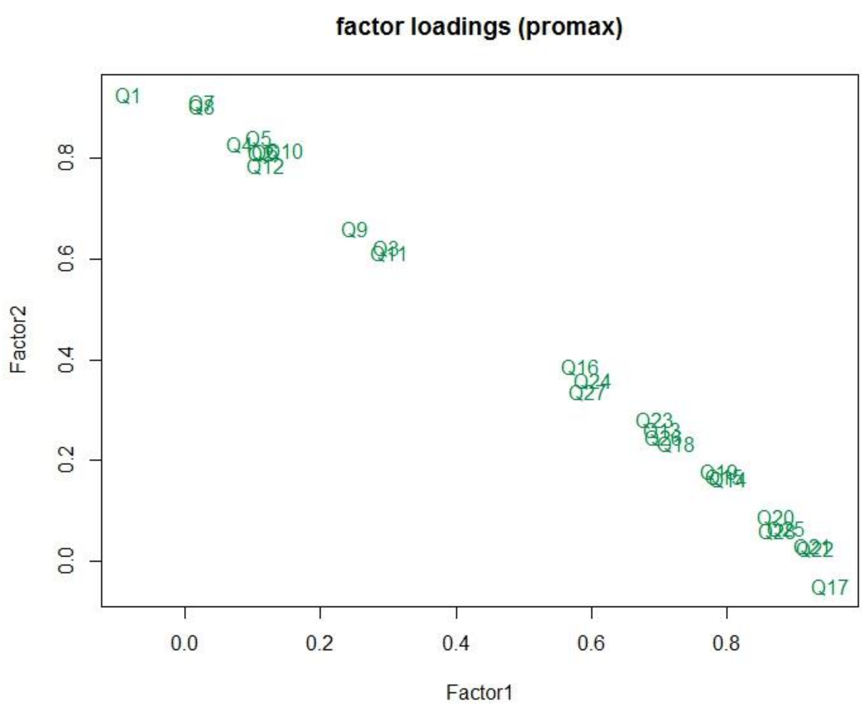


Figure 10(b) FA results (promax)

```
Loadings:
    Factor1 Factor2
Q1          0.927
Q2   0.114  0.811
Q3   0.299  0.622
Q4          0.829
Q5   0.110  0.841
Q6   0.120  0.812
Q7          0.912
Q8          0.905
Q9   0.252  0.661
Q10  0.147  0.817
Q11  0.302  0.614
Q12  0.119  0.786
Q13  0.704  0.262
Q14  0.802  0.164
Q15  0.795  0.167
Q16  0.583  0.387
Q17  0.952
Q18  0.725  0.234
Q19  0.788  0.178
Q20  0.872
Q21  0.926
Q22  0.929
Q23  0.692  0.280
Q24  0.602  0.360
Q25  0.887
Q26  0.706  0.246
Q27  0.594  0.337
Q28  0.874

               Factor1 Factor2
SS loadings     10.221   8.471
Proportion Var   0.365   0.303
Cumulative Var   0.365   0.668

Factor Correlations:
        Factor1 Factor2
Factor1   1.000   0.826
Factor2   0.826   1.000
```

```
Loadings:
    Factor1 Factor2 Factor3
Q1          0.935
Q2   0.100  0.821
Q3   0.282  0.633
Q4          0.838
Q5          0.852
Q6   0.105  0.822
Q7          0.923
Q8          0.917
Q9   0.237  0.673
Q10  0.134  0.829
Q11  0.286  0.625
Q12  0.108  0.798
Q13  0.677  0.267   0.231
Q14  0.780  0.163   0.259
Q15  0.770  0.173   0.213
Q16  0.561  0.399   0.110
Q17  0.926          0.158
Q18  0.701  0.249   0.116
Q19  0.766  0.196
Q20  0.853  0.107
Q21  0.914
Q22  0.921
Q23  0.688  0.296
Q24  0.598  0.376
Q25  0.877
Q26  0.700  0.261
Q27  0.593  0.353
Q28  0.863

               Factor1 Factor2 Factor3
SS loadings      9.799   8.753   0.254
Proportion Var   0.350   0.313   0.009
Cumulative Var   0.350   0.663   0.672

Factor Correlations:
        Factor1 Factor2 Factor3
Factor1  1.0000   0.821  0.0331
Factor2  0.8215   1.000  0.1008
Factor3  0.0331   0.101  1.0000
```

I use both orthogonal rotation (varimax) and oblique rotation (promax) for my Factor Analysis. Although the tests provided by MLE approach for estimation give me really small p-values, I still use two to three factors to do my Factor Analysis. By applying Factor Analysis, I notice that in fact, two-factor results sufficiently give me reasonable explanations (cumulative variances are large enough for two-factor results). From the results obtained by orthogonal rotation (independent factors), I see factor 1 represents the evaluations related to instructors (Q13 – Q28) and factor 2 represents the evaluations related to courses (Q1 – Q12). This result is clearly exhibited in the plot of factor loadings. From the results obtained by oblique rotation (correlated factors), I see even more clearly factor 1 represents the evaluations related to instructors (Q13 – Q28) and factor 2 represents the evaluations related to courses (Q1 – Q12). Also, I observe that the correlation between factor 1 and factor 2 is strong (0.826). Therefore, I need to further analyze the dependency pattern.

CANONICAL CORRELATION ANALYSIS (CCA)

In order to analyze the dependency pattern between two sets of data, I next apply Canonical Correlation Analysis to the data set. The two sets of data are: (1) scores of evaluations related to instructors (Q13 – Q28) and (2) scores of evaluations related to courses (Q1 – Q12). The results are shown below in Figure 11, Figure 12, Figure 13 and Figure 14,

Figure 11 correlation between canonical covariates

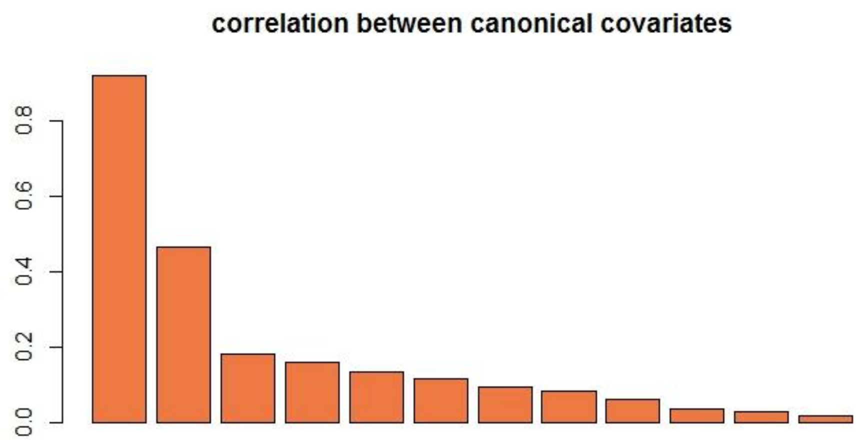


Figure 12(a) correlation between F1 and G1
Figure 12(b) correlation between F2 and G2

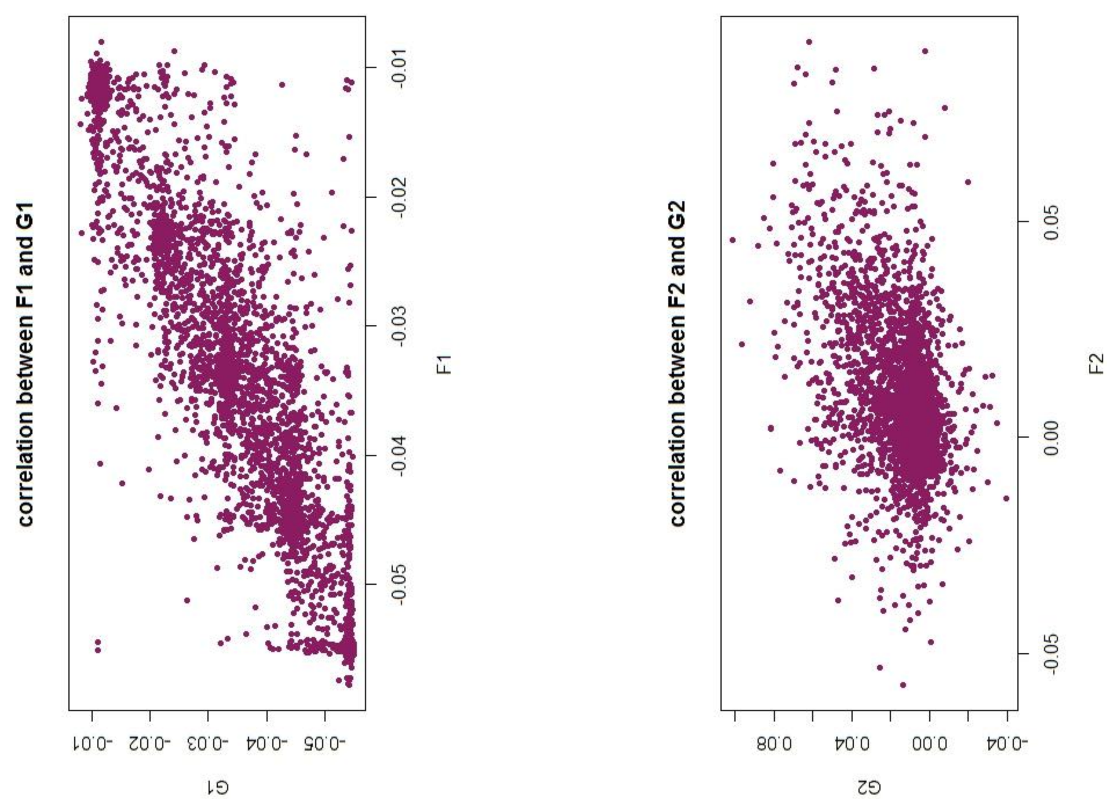


Figure 13 test for dimension of canonical covariates

```
Wilks' Lambda, using F-approximation (Rao's F):
              stat     approx df1      df2      p.value
1 to 12:   0.1059784 76.2572932 192 55976.73 0.000000e+00
2 to 12:   0.6980421 12.8834933 165 51779.35 0.000000e+00
3 to 12:   0.8919440  4.7673849 140 47563.35 0.000000e+00
4 to 12:   0.9223553  4.0262207 117 43323.96 0.000000e+00
5 to 12:   0.9470721  3.2973467  96 39054.53 0.000000e+00
6 to 12:   0.9642340  2.7509402  77 34745.50 1.798561e-14
7 to 12:   0.9777429  2.1801149  60 30382.58 3.686516e-07
8 to 12:   0.9866076  1.7403230  45 25943.44 1.548541e-03
9 to 12:   0.9934886  1.1851827  32 21390.94 2.174450e-01
10 to 12:  0.9975230  0.6854023  21 16657.89 8.520219e-01
11 to 12:  0.9988452  0.5588153  12 11604.00 8.763698e-01
12 to 12:  0.9996747  0.3776210   5  5803.00 8.643768e-01
```

Figure 14 x coefficients and y coefficients

[,1]	[,2]
3.143437e-05	-2.746157e-04
-4.852567e-05	6.232623e-05
-7.562944e-05	4.721947e-04
-2.466570e-05	1.771631e-04
-9.443944e-05	-2.382250e-04
-4.498103e-05	1.712598e-04
-1.177169e-05	-1.736988e-04
-2.889232e-05	-3.286529e-04
-4.259165e-05	3.772787e-04
-1.179894e-04	-3.022804e-04
-5.050612e-05	4.906914e-04
-5.463175e-05	-3.917900e-04

[,1]	[,2]
-9.033718e-05	-1.841185e-04
-3.424820e-05	1.174338e-04
-2.909066e-05	1.865809e-04
-1.164478e-04	-3.241919e-04
9.251931e-06	4.763511e-04
-9.203138e-06	-1.275292e-04
-5.484755e-05	-1.408147e-05
-1.157290e-05	-2.208361e-05
6.165087e-06	2.836094e-04
5.887491e-06	1.813705e-04
-5.301290e-05	-2.851844e-04
-8.747940e-05	-3.442080e-04
2.700867e-05	1.982076e-04
-6.204237e-05	-1.785516e-05
-6.563837e-05	-2.083470e-04
1.064056e-05	1.833497e-04

In Canonical Correlation Analysis, I find there are in total 12 canonical correlations, which means the dimension of canonical covariates is 12. After applying the asymptotic tests for the statistical significance of canonical correlation coefficients, I notice that only the p-values for the first 8 canonical correlations are pretty small (<0.05), so I can say that the first 8 dimensions of canonical covariates are statistically significant. Checking the bar plot, I observe that the first 2 pairs of canonical covariates are highly correlated. In order to get more evidence, I apply the scatterplots for the first two canonical correlations and I find the results I get from the bar plot are generally correct. According to the x coefficients and y coefficients, I discover F1 mainly represents Q10: “My initial expectations about the course were met at the end of the period or year", G1 mainly represents Q16: "The Instructor was committed to the course and was understandable" (Q10 and Q16 are positively correlated), F2 mainly represents Q11: "The course was relevant and helpful to my professional development" and G2 mainly represents Q17: "The Instructor arrived on time for classes" (Q11 and Q17 are positively correlated).

Problem #2

DATA PREPROCESSING

The big data set that is provided contains 14 small data sets. In order to effectively analyze data, I need to combine the 14 small data sets together in a correct manner. First, I sort each small data set by its row names. I find the row names for raw files all start with 5 and the row names for processed files all start with 1. Therefore, I treat the row names for raw files as "original names – 4" and then combine each raw file and processed file by column. Next, I combine each combined file from last step by row to produce a big data set as a whole with rows all observations and columns all variables, delete the irrelevant column "timestamp" and rename each row in a manner of "instance_number", such as "a1_1".

BASIC EXPLORATORY DATA ANALYSIS

After preprocessing the data set, I do some basic explorations. For example, I check the dimensionality, which is 9873*50; I use the function summary() to get a summary of the data set; I check the variance of each variable; I plot the correlation matrix; and I check the outliers. The results are shown below in Figure 15, Figure 16, and Figure 17,

Figure 15 correlation matrices

Figure 16 adjusted quantile for outliers

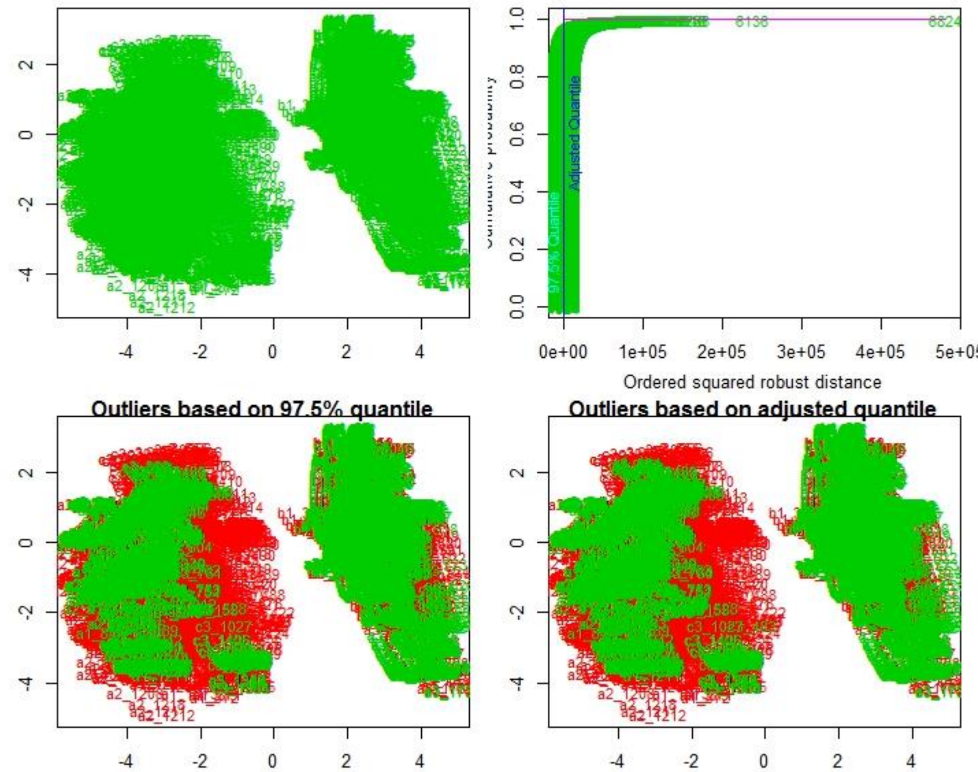


Figure 17 variance for each variable

```
          lhx           lhy           lhz           rhx           rhy           rhz            hx
1.856693e+00 6.714363e-01 1.137048e-01 1.097646e+00 1.270121e+00 1.206233e-01 1.281855e+00
           hy            hz            sx            sy            sz           lwx           lwy
1.383589e+00 9.571164e-02 1.314601e+00 2.006359e-01 1.033809e-01 1.744064e+00 5.075862e-01
          lwz           rwx           rwy           rwz            x1            x2            x3
1.121876e-01 1.067893e+00 1.010872e+00 1.230202e-01 5.159848e-05 9.634730e-05 5.015779e-05
           x4            x5            x6            x7            x8            x9           x10
8.078124e-05 1.639124e-04 5.714781e-05 3.591332e-05 6.726513e-05 3.589018e-05 5.159298e-05
          x11           x12           x13           x14           x15           x16           x17
1.124362e-04 4.118850e-05 2.674488e-06 3.710166e-06 1.627197e-06 4.253944e-06 5.025048e-06
          x18           x19           x20           x21           x22           x23           x24
1.016665e-06 1.876260e-06 2.657086e-06 1.281010e-06 2.577227e-06 3.428373e-06 9.100096e-07
          x25           x26           x27           x28           x29           x30           x31
1.247504e-04 1.782549e-04 8.879830e-05 1.227091e-04 5.827027e-06 7.281548e-06 4.191807e-06
          x32
4.797050e-06
```

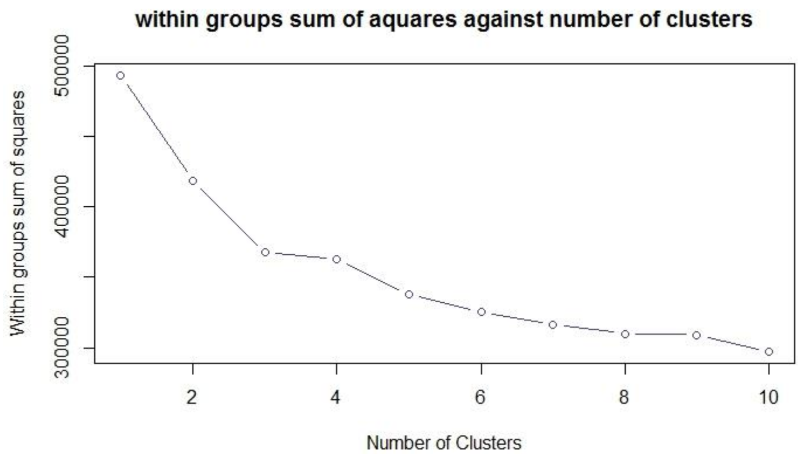
Since from Figure 17, I observe that variances for variables vary greatly, I use 3 methods to standardize the data set (these 3 methods can be found in R file). Then, I check the correlation matrices (Figure 15), and find they are all the same, which means the standardizations do not change the correlation matrix. From the correlation matrix, I observe some interesting results. For example, x-coordinates are correlated with x-coordinates, y-coordinates are correlated with y-coordinates, and z-coordinates are correlated with z-coordinates. I also use the function aq.plot() to check the outliers. From Figure 17, I notice there are many obscure outliers. Therefore, I decide to ignore the outlier checking process and start to do clustering.

K-MEANS CLUSTERING

First, I will try k-means clustering. Before applying k-means clustering, I need to check how many clusters I need for clustering, i.e. what the number of k is. By plotting the plot of within groups sum of squares against number of clusters (Figure 18), I notice it is reasonable to choose

3 clusters since the “elbow” is at cluster 3, but I will try 3, 4 and 5 clusters in the analysis below.

Figure 18 within groups sum of squares against number of clusters



In k-means clustering, I first try 3-means clustering, as Figure 18 suggests. After applying 3-means clustering, in order to visualize the clusters, I need to plot the observations on the plot of PC 1 and PC 2. It is clear that PC 1 separates the <u>cluster 1 (palegreen4)</u> from the cluster2 (royalblue3) and PC 2 separates the <u>cluster 3 (indianred 2)</u> from the cluster 1 and cluster 2 (Figure 20).

Figure 19(a) PCA results

```
> summary(ges.pc)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4    Comp.5     Comp.6     Comp.7
Standard deviation     3.0361226 2.4104878 1.95483960 1.73890163 1.6431905 1.58561357 1.47328406
Proportion of Variance 0.1843608 0.1162090 0.07642796 0.06047558 0.0540015 0.05028341 0.04341132
Cumulative Proportion  0.1843608 0.3005698 0.37699780 0.43747337 0.4914749 0.54175828 0.58516960
                           Comp.8     Comp.9    Comp.10    Comp.11    Comp.12    Comp.13    Comp.14
Standard deviation     1.42161774 1.39615710 1.28174704 1.26688539 1.20705752 1.19671438 1.09966945
Proportion of Variance 0.04041994 0.03898509 0.03285751 0.03209997 0.02913976 0.02864251 0.02418546
Cumulative Proportion  0.62558954 0.66457463 0.69743214 0.72953211 0.75867187 0.78731438 0.81149983
                          Comp.15    Comp.16    Comp.17    Comp.18    Comp.19    Comp.20    Comp.21
Standard deviation     1.06689422 1.01550032 0.97949892 0.91808275 0.91595495 0.78317740 0.71212314
Proportion of Variance 0.02276527 0.02062482 0.01918836 0.01685752 0.01677947 0.01226734 0.01014239
Cumulative Proportion  0.83426510 0.85488992 0.87407828 0.89093580 0.90771527 0.91998261 0.93012499
                           Comp.22     Comp.23     Comp.24     Comp.25     Comp.26     Comp.27
Standard deviation     0.680307553 0.624652342 0.597788364 0.541735665 0.512746722 0.491578300
Proportion of Variance 0.009256367 0.007803811 0.007147019 0.005869551 0.005258184 0.004832984
Cumulative Proportion  0.939381360 0.947185171 0.954332190 0.960201741 0.965459925 0.970292909
                           Comp.28     Comp.29     Comp.30     Comp.31     Comp.32     Comp.33
Standard deviation     0.490070653 0.472958652 0.458460201 0.315771461 0.306883712 0.297629530
Proportion of Variance 0.004803385 0.004473798 0.004203715 0.001994232 0.001883552 0.001771667
Cumulative Proportion  0.975096294 0.979570092 0.983773807 0.985768039 0.987651591 0.989423258
                           Comp.34     Comp.35     Comp.36    Comp.37     Comp.38      Comp.39
Standard deviation     0.294556759 0.272974748 0.262405033 0.24976187 0.239606677 0.2104549910
Proportion of Variance 0.001735274 0.001490304 0.001377128 0.00124762 0.001148227 0.0008858261
Cumulative Proportion  0.991158532 0.992648836 0.994025964 0.99527358 0.996421811 0.9973076372
                            Comp.40      Comp.41      Comp.42      Comp.43      Comp.44      Comp.4
5
Standard deviation     0.2012913068 0.1943608190 0.1578140190 0.0831692722 0.0797585887 0.070200819
2
Proportion of Variance 0.0008103638 0.0007555226 0.0004981053 0.0001383426 0.0001272286 0.000098563
1
Cumulative Proportion  0.9981180010 0.9988735236 0.9993716289 0.9995099714 0.9996372001 0.999735763
2
                            Comp.46      Comp.47      Comp.48      Comp.49      Comp.50
Standard deviation     6.251441e-02 5.433706e-02 4.976758e-02 4.861677e-02 3.886992e-02
Proportion of Variance 7.816103e-05 5.905033e-05 4.953624e-05 4.727181e-05 3.021741e-05
Cumulative Proportion  9.998139e-01 9.998730e-01 9.999225e-01 9.999698e-01 1.000000e+00
```

I notice from Figure 19(a) that I actually need 16 principal components in order to make cumulative proportion greater than 85%. However, in 3-means clustering, only PC 1 and PC 2 are crucial for data separations. I observe from Figure 19(b) that PC 1 represents the changes of positions of left-right movements and PC 2 represents the changes of positions of front-back movements.

Figure 19(b) PCA results

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
lhx	-0.266	-0.118	-0.107						
lhy		-0.212	-0.256						
lhz	0.268	-0.218							
rhx	-0.266								
rhy		-0.230	-0.241						
rhz	0.272	-0.200							
hx	-0.300	-0.113	-0.101						
hy		-0.372							
hz	0.264	-0.224							
sx	-0.301	-0.110							
sy	-0.139	-0.319	-0.108						
sz	0.271	-0.218							
lwx	-0.269	-0.136	-0.105						
lwy		-0.241	-0.258						
lwz	0.269	-0.220							
rwx	-0.285								
rwy		-0.245	-0.238						
rwz	0.275	-0.200							
x1			0.104	-0.309	0.139	0.116	-0.292	0.145	-0.192
x2					-0.211	0.461	0.139		
x3				0.115	-0.357			0.256	-0.256
x4			-0.120	0.340				-0.266	-0.271
x5				-0.264	-0.272			-0.351	
x6					-0.274		-0.348		0.294
x7			0.105	-0.314	0.145	0.100	-0.289	0.137	-0.182
x8					-0.208	0.457	0.142		
x9				0.120	-0.361			0.259	-0.266
x10			-0.118	0.341				-0.282	-0.263
x11				-0.260	-0.271			-0.359	
x12					-0.280		-0.351		0.299
x13				-0.181	0.124	0.192	-0.226		-0.236
x14						0.419	0.160		0.142
x15					-0.293			0.311	-0.155
x16				0.246		0.190	-0.195	-0.144	-0.170
x17				-0.241	-0.136	-0.164	0.177	-0.250	-0.223
x18					-0.178		-0.269		0.183
x19				-0.185	0.131	0.164	-0.242		-0.241
x20						0.417	0.173		0.152
x21					-0.278			0.316	-0.158
x22				0.245		0.167	-0.197	-0.168	-0.174
x23				-0.233	-0.138	-0.149	0.162	-0.266	-0.230
x24					-0.181		-0.262		0.198
x25		-0.188	0.296	0.132			-0.105		
x26		-0.188	0.288				0.112		
x27		-0.187	0.296	0.128			-0.106		
x28		-0.192	0.288				0.107		
x29	-0.103	-0.129	0.248	0.122					
x30	-0.109	-0.104	0.268				0.124		
x31	-0.108	-0.131	0.240	0.116					
x32	-0.108	-0.127	0.266				0.119		

In k-means clustering, I then try 4-means clustering. According to Figure 21(a) and Figure 21(b), I find that PC 1 separates cluster 3 from cluster 1 and cluster 2, PC 2 separates cluster 4 (khaki3) from cluster1, cluster 2 and cluster 3, and PC 3 separates cluster 1 from cluster 4. From Figure 19(b), I notice that PC 3 represents the parallel changes of positions of left-right movements.

Figure 20 3-means clustering

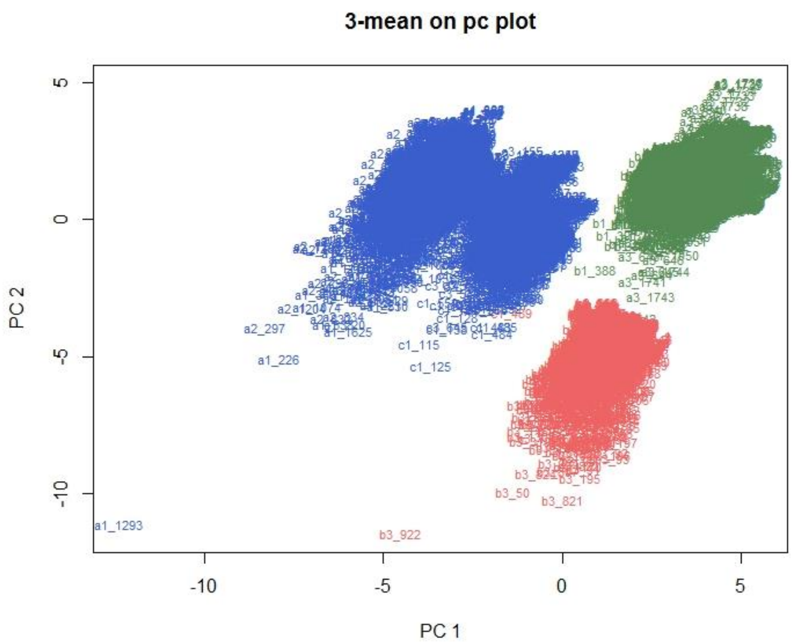


Figure 21(a) 4-means clustering

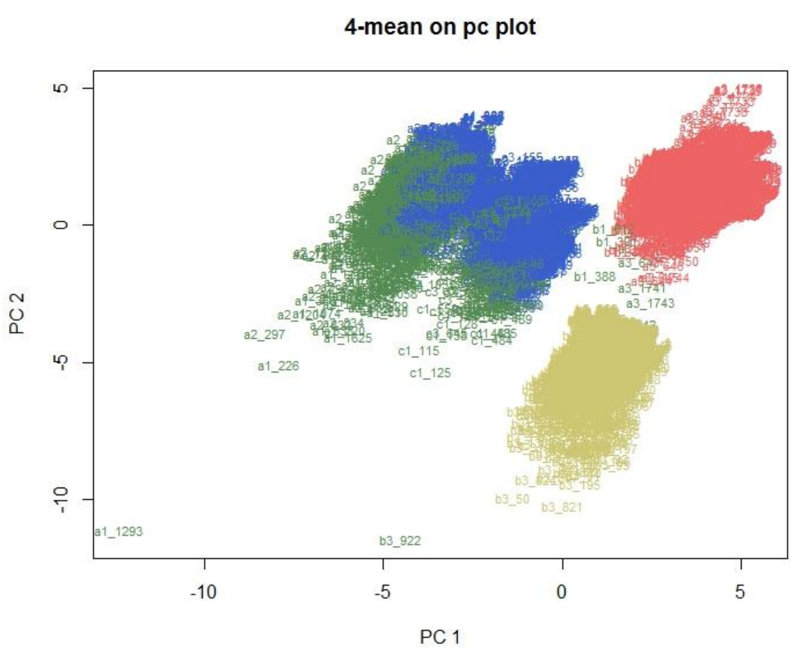
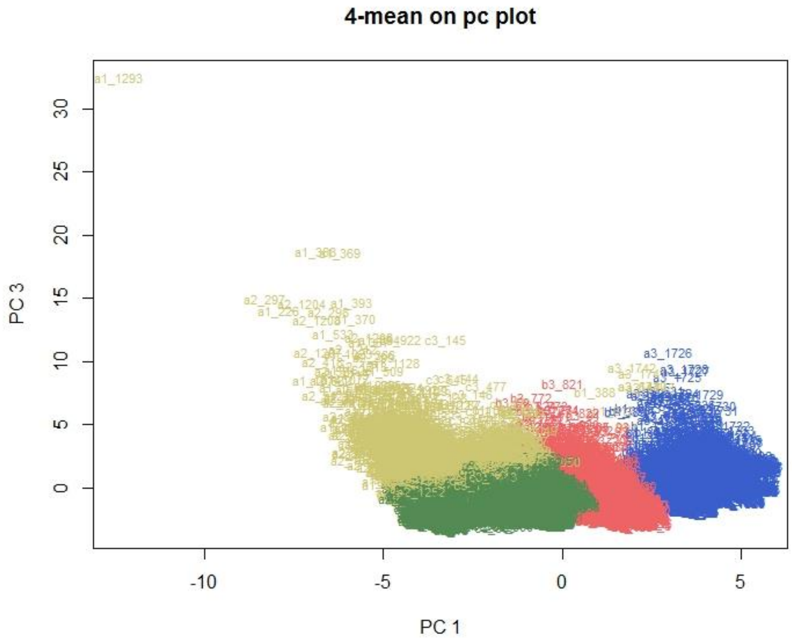


Figure 21(b) 4-means clustering



In k-means clustering, I finally try 5-means clustering. According to Figure 22(a), Figure 22(b) and Figure 22(c), I find that PC 1 separates cluster 2 from cluster 1 and cluster 3, PC 2 roughly separates cluster 5 (darkorchid3) from cluster 4, PC 3 separates cluster 1 from cluster 4, and PC 5 separates cluster 3 from cluster 1. From Figure 19(b), I notice that PC 5 represents the velocities of left-right movements.

Figure 22(a) 5-means clustering

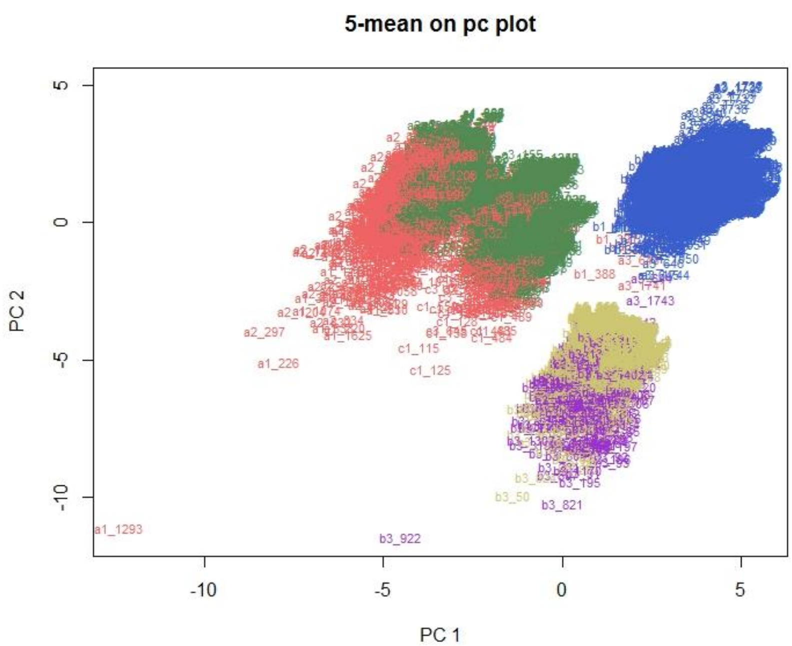


Figure 22(b) 5-means clustering

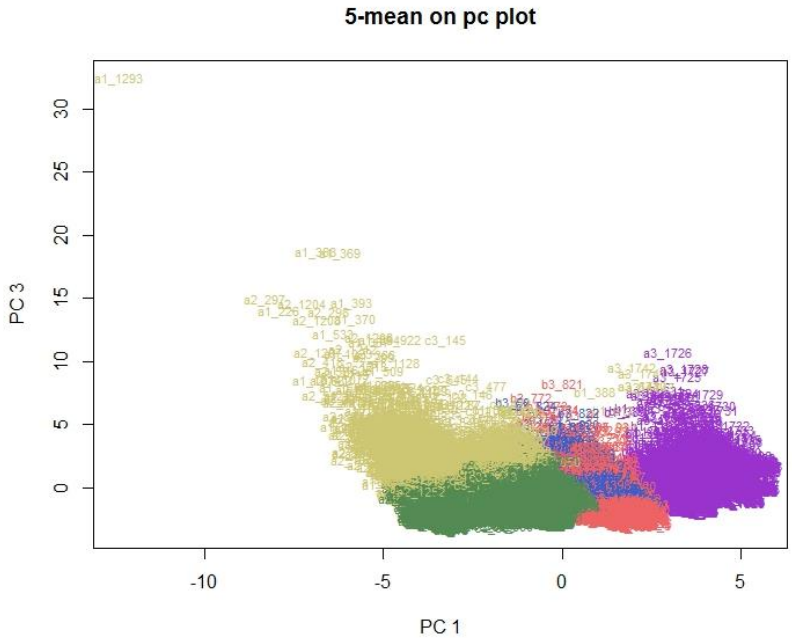
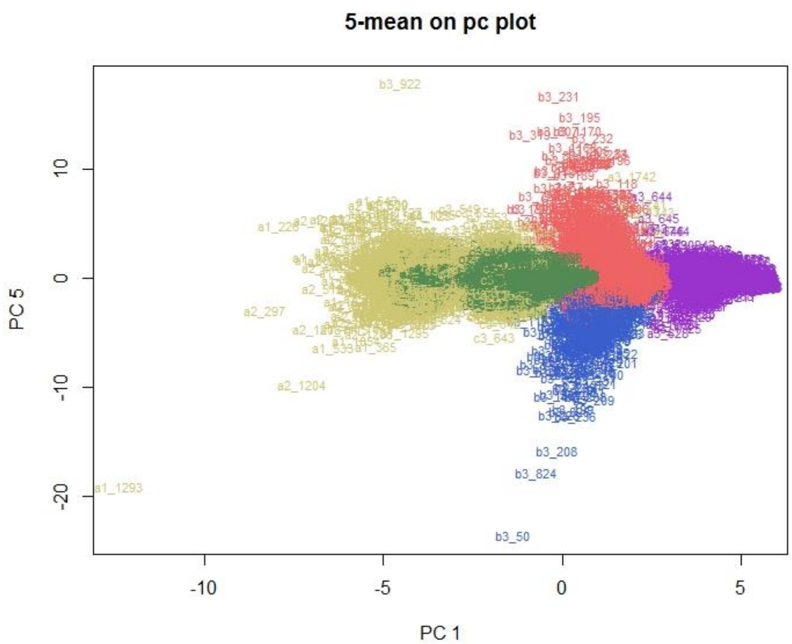


Figure 22(c) 5-means clustering



MODEL-BASED CLUSTERING

Next, I will try model-based clustering. By applying the function Mclust(), I am suggested to use VVV (ellipsoidal, varying volume, shape, and orientation) model with 9 clusters, as shown in Figure 23(a) and 23(b). I find from the clustering table that cluster 3 (color: firebrick1), cluster 4 (color: goldenrod3), cluster 6 (color: lightgreen) and cluster 8 (color: orange1) are large clusters. If we only consider the large clusters, the separation manner of model-based clustering is similar to the separation manner of 4-means clustering (Figure 24).

Figure 23(a) summary of model-based clustering

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VVV (ellipsoidal, varying volume, shape, and orientation) model with 9 components:

 log.likelihood    n    df      BIC      ICL
       92647.01 9873 11933 75539.54 75468.32

Clustering table:
   1    2    3    4    5    6    7    8    9
 977  666 1895 1746  553 1863  138 1419  616
```

Figure 23(a) summary of model-based clustering

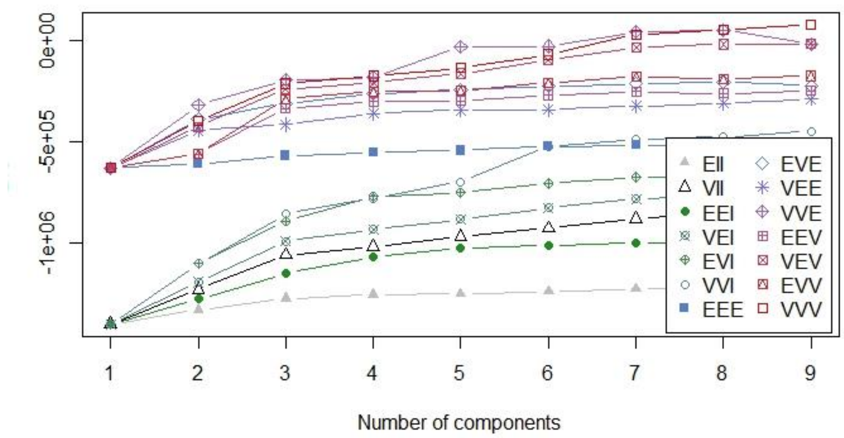
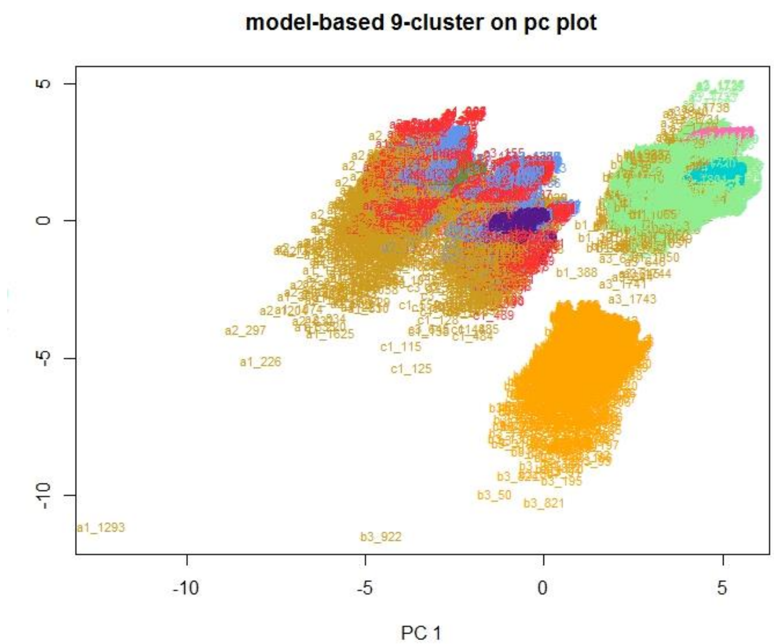


Figure 24 model-based clustering



HIERARCHICAL CLUSTERING

Finally, I will try hierarchical clustering. I observe from Figure 25, Figure 26 and Figure 27 that hierarchical clustering does not do a good job. I find that even though I set k = 9, i.e. 9 clusters, one cluster has huge number of observations while others only have small number of observations. I believe the reason is that there are many outliers existing in the data set, which is shown in Figure 16. Therefore, we had better adopt the results from k-means clustering and model-based clustering.

Figure 25 hierarchical clustering (complete-linkage)

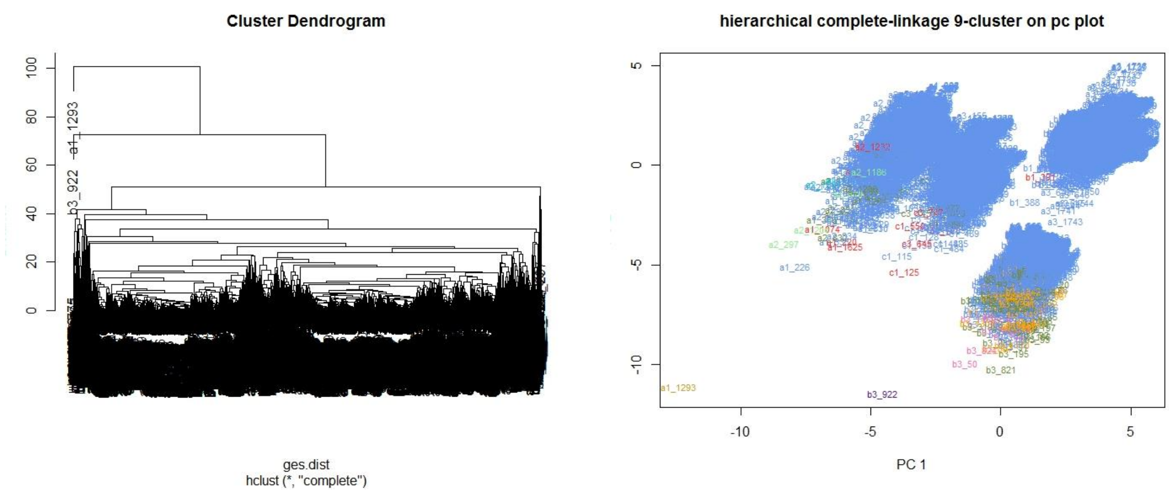


Figure 26 hierarchical clustering (average-linkage)

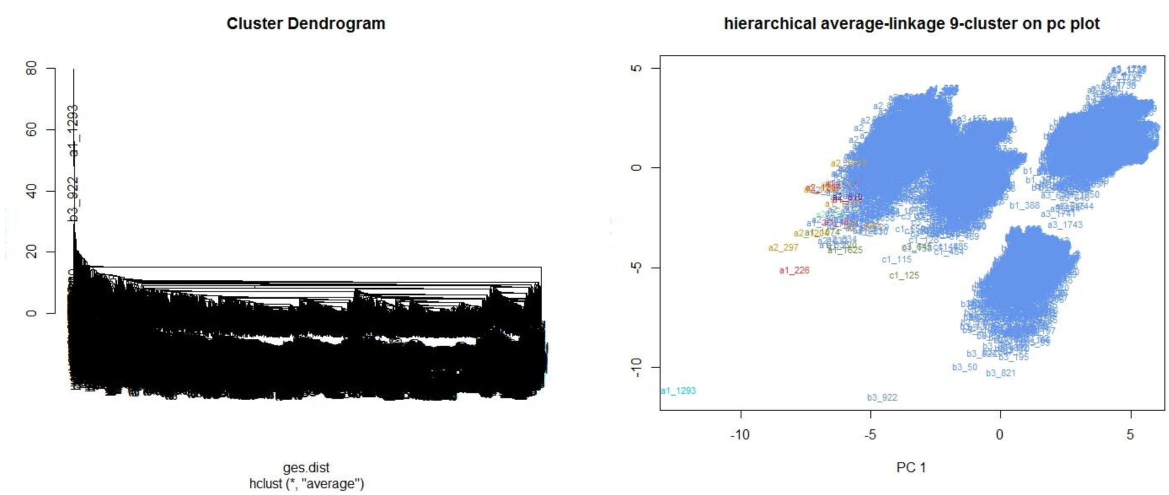
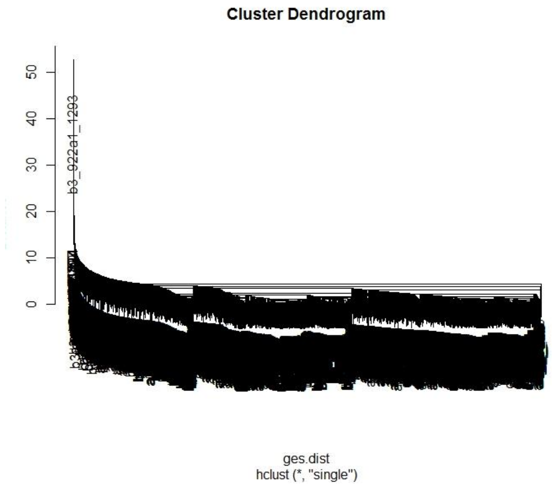
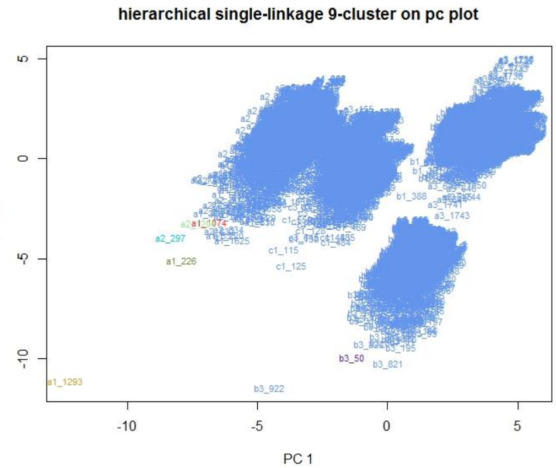


Figure 27 hierarchical clustering (single-linkage)


Cluster Dendrogram
50
40
30
20
10
0
b3_922a1_1293
ges.dist
hclust (*, "single")


hierarchical single-linkage 9-cluster on pc plot
5
0
-5
-10
a1_1293
b3_922
a2_297
a1_226
c1_115
c1_125
b1_388
a3_1743
b3_50
b3_821
-10
-5
0
5
PC 1