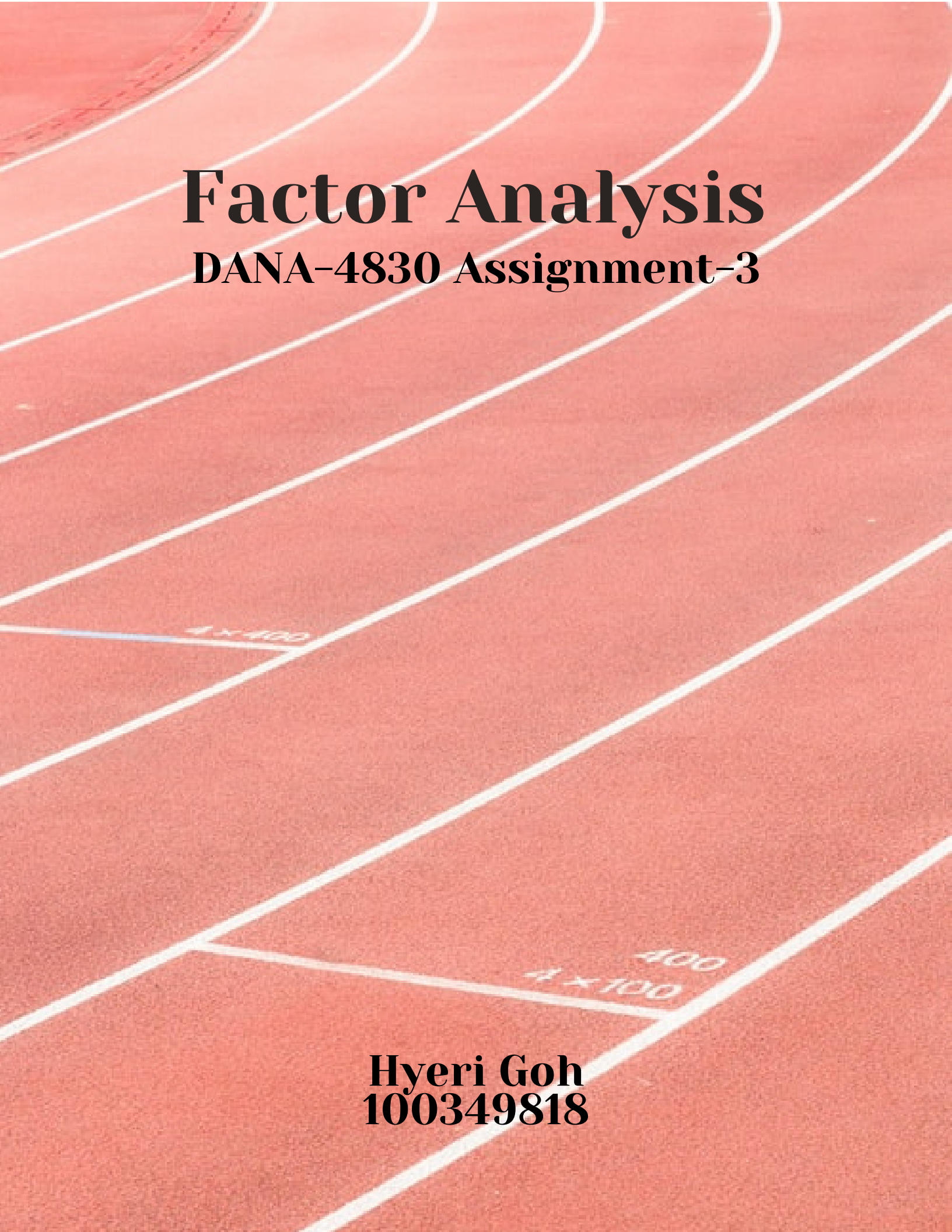


Factor Analysis

DANA-4830 Assignment-3

Hyeri Goh
100349818

Accuracy & Missing value

Since the survey is using Likert scale which has seven levels as -3 for 'strongly disagree', 0 for 'neutral', and +3 for 'strongly agree', all observations should fall into the range of (-3, 3) and has integer value. All observations under each 25 columns were in the range of (-3, 3), however, there were 4 columns that include non-integer value. (Figure 1, Figure 2)

```
> summary(data)
       Q1              Q2              Q3              Q4              Q5              Q6              Q7              Q8              Q9       
 Min.   :-3.00   Min.   :-3.000   Min.   :-3.0000   Min.   :-3.000   Min.   :-3.000   Min.   :-3.000   Min.   :-3.000   Min.   :-3.000   Min.   :-3.000  
 1st Qu.: 2.00   1st Qu.: 2.000   1st Qu.: 0.0000   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 1.000  
 Median : 3.00   Median : 3.000   Median : 1.0000   Median : 3.000   Median : 2.000   Median : 2.000   Median : 3.000   Median : 3.000   Median : 3.000  
 Mean   : 2.38   Mean   : 2.509   Mean   : 0.6751   Mean   : 2.312   Mean   : 1.916   Mean   : 1.924   Mean   : 2.529   Mean   : 2.393   Mean   : 1.921  
 3rd Qu.: 3.00   3rd Qu.: 3.000   3rd Qu.: 2.0000   3rd Qu.: 3.000   3rd Qu.: 3.000   3rd Qu.: 3.000   3rd Qu.: 3.000   3rd Qu.: 3.000   3rd Qu.: 3.000  
 Max.   : 3.00   Max.   : 3.000   Max.   : 3.0000   Max.   : 3.000   Max.   : 3.000   Max.   : 3.000   Max.   : 3.000   Max.   : 3.000   Max.   : 3.000  
      Q10             Q11             Q12             Q13             Q14             Q15             Q16             Q17             Q18       
 Min.   :-3.000   Min.   :-3.000   Min.   :-1.000   Min.   :-3.000   Min.   :-3.000   Min.   :-3.000   Min.   :-3.0000   Min.   :-3.0000   Min.   :-3.000  
 1st Qu.: 2.000   1st Qu.: 3.000   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.:-3.000   1st Qu.: 1.000   1st Qu.:-2.0000   1st Qu.:-1.0000   1st Qu.: 2.000  
 Median : 2.000   Median : 3.000   Median : 3.000   Median : 2.000   Median :-2.000   Median : 2.000   Median : 0.0000   Median : 1.0000   Median : 3.000  
 Mean   : 2.171   Mean   : 2.671   Mean   : 2.603   Mean   : 1.603   Mean   :-1.586   Mean   : 1.753   Mean   :-0.1033   Mean   : 0.8035   Mean   : 2.379  
 3rd Qu.: 3.000   3rd Qu.: 3.000   3rd Qu.: 3.000   3rd Qu.: 3.000   3rd Qu.: 0.000   3rd Qu.: 3.000   3rd Qu.: 1.0000   3rd Qu.: 2.0000   3rd Qu.: 3.000  
 Max.   : 3.000   Max.   : 3.000   Max.   : 3.000   Max.   : 3.000   Max.   : 3.000   Max.   : 3.000   Max.   : 3.0000   Max.   : 3.0000   Max.   : 3.000  
      Q19             Q20             Q21             Q22             Q23             Q24             Q25       
 Min.   :-3.000   Min.   :-3.000   Min.   :-3.000   Min.   :-3.0000   Min.   :-3.000   Min.   :-3.000   Min.   :-3.000  
 1st Qu.: 2.000   1st Qu.: 0.000   1st Qu.:-3.000   1st Qu.:-2.0000   1st Qu.:-3.000   1st Qu.: 2.000   1st Qu.: 2.000  
 Median : 3.000   Median : 2.000   Median :-3.000   Median : 0.0000   Median :-3.000   Median : 3.000   Median : 3.000  
 Mean   : 2.222   Mean   : 1.174   Mean   :-2.208   Mean   :-0.4779   Mean   :-2.169   Mean   : 2.295   Mean   : 2.127  
 3rd Qu.: 3.000   3rd Qu.: 3.000   3rd Qu.:-2.000   3rd Qu.: 1.0000   3rd Qu.:-2.000   3rd Qu.: 3.000   3rd Qu.: 3.000  
 Max.   : 3.000   Max.   : 3.000   Max.   : 3.000   Max.   : 3.0000   Max.   : 3.000   Max.   : 3.000   Max.   : 3.000  
```

Figure 1

```
> str(data)
'data.frame'
 $ Q1 : int
 $ Q2 : int
 $ Q3 : int
 $ Q4 : int
 $ Q5 : num
 $ Q6 : int
 $ Q7 : int
 $ Q8 : int
 $ Q9 : int
 $ Q10: num
 $ Q11: int
 $ Q12: int
 $ Q13: int
 $ Q14: int
 $ Q15: int
 $ Q16: int
 $ Q17: int
 $ Q18: int
 $ Q19: int
 $ Q20: num
 $ Q21: int
 $ Q22: num
 $ Q23: int
 $ Q24: int
 $ Q25: int
```

Q5	-3 4	-2 10	-1 16	0 74	1 128	1.915510719 1	2 244	3 317
Q10	-3 3	-2 6	-1 13	0 36	1 96	2 256	2.170670038 3	3 381
Q20	-3 38	-2 42	-1 65	0 87	1 155	1.174022699 1	2 179	3 227
Q22	-3 182	-2 114	-1 99	-0.477931904 1	0 112	1 137	2 90	3 59

Figure 2

In order to check where decimal values come from, I calculated the mean value of those 4 columns – Q5, Q10, Q20, Q22 – excluding the cells with decimal values. It turned out the mean of 4 columns and the decimal values were exactly match each other, which implies the observations with decimal value had been missing at the first time and then imputed with the mean value of corresponding columns later. It also explains why the given dataset does not include any blank cell. Even though the decimal values are against the range of Likert scale, because they are derived from the mean value of data collected, decimal values will be kept.

Outliers

For detecting outliers, Mahalanobis distance and cutoff point were calculated. Since there are 25 columns which will be subject to factor analysis, degrees of freedom is 24 (=25-1) and cutoff point is 52.61966. After comparing Mahalanobis distance and cutoff point, there were 69 outliers whose Mahalanobis distance is greater than cutoff point. (Figure 3) These 69 observations were removed, which left 725 rows in clean dataset named 'nooutlier'.

```
> cutoff=qchisq(1-.001, ncol(data)) #df = 25-1 = 24
> cutoff
[1] 52.61966
> table(mahal < cutoff)

FALSE  TRUE
   69   725
```

Figure 3

Assumptions Check

Before checking assumptions, I generated 725 random numeric values that follow chi-squared distribution, then built linear model that has random values as response variable and 25 columns (Q1 ~ Q25 from 'nooutlier' dataframe) as predictors. From linear model, I plotted the studentized residuals.

1. Normality
   According to the histogram of studentized residuals, it shows slightly right-skewed shape rather than normal bell shape. (Figure 4) I also conducted shapiro test to check the assumption which resulted in very small p-value. It also indicates variables are significantly different from normal distribution.

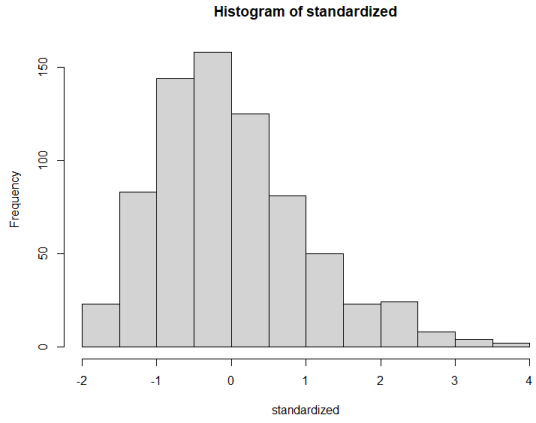


Figure 4

2. Homoscedasticity/Heteroscedasticity
   Studentized residual plot was used to test whether the predictors – columns Q1 ~ Q25 – have equal or non-equal variances. The plot below indicates that variance is constant for all settings of the predictors because there is no pattern or trend shown. (Figure 5) Hence, Homoscedasticity assumption is valid.

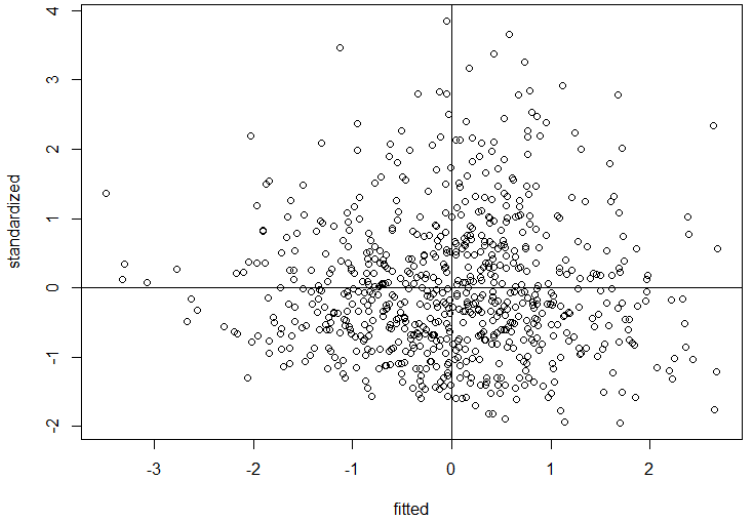


Figure 5

3. Linearity
   QQ-plot of studentized residuals shows a few of studentized residuals affect to non-linearity since there are some gaps between QQ line and observations. (Figure 6) Also, from studentized residual plot (Figure 5), there are several residuals that are mapped out of the range [-2, 2]. Considering that good residual plots should show 95% of residuals within [-2, 2], linearity assumption is not met.

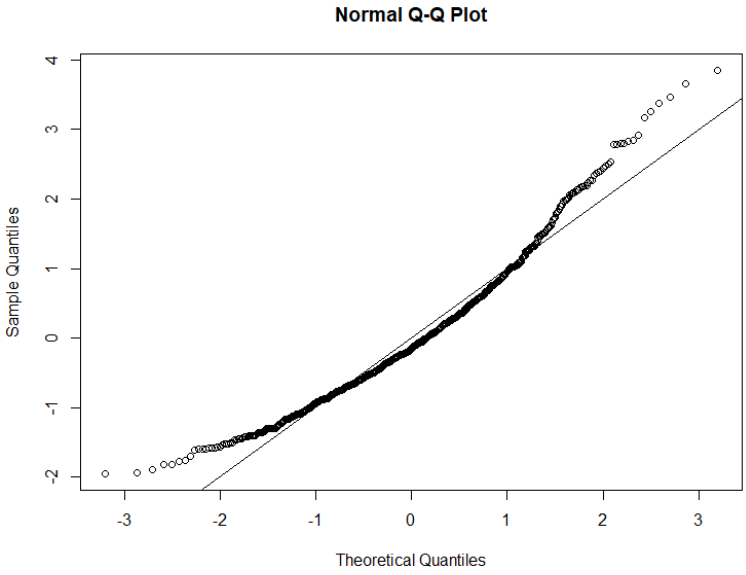


Figure 6

4. Additivity
   Before factor analysis, it should be checked if variables are correlated or not. Correlation matrix shows there are several combinations of variables with moderate or high correlation. Therefore, additivity is valid.

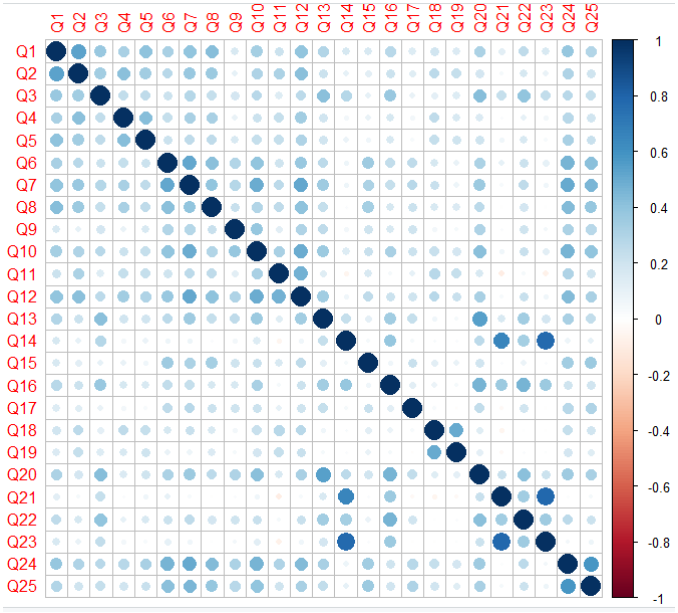


Figure 7

5. Bartlett's test and KMO
   Bartlett's test proved that factor analysis might be useful since the p-value is very small. Also, KMO test showed overall MSA = 0.89. Considering the closer to 1 overall MSA is, the better, it can be concluded that there are a significant number of factors in the dataset.

After assumption check, it was turned out some of assumptions are not met. However, because there are obvious correlations among items, it is better to conduct factor analysis and find out proper interpretation to explain the survey result.

Number of factors

The Kaiser criterion recommend how many factors should be used through eigenvalues. (Figure 8) The Kaiser criterion with the old rule that extracts the number of eigenvalues over 1 suggested 2 factors, while the one with the new rule that gives the number of eigenvalues over 0.7 suggested 3 factors.

```
> no.factors$fa.values #eigenvalues
 [1]  6.129755778  2.101307741  0.993192099  0.464913272  0.391574822  0.209136285  0.088211671  0.036990469  0.008570025 -0.010057949
[11] -0.076400745 -0.090318852 -0.110989791 -0.129734716 -0.154132425 -0.175797284 -0.213357345 -0.250335084 -0.270358261 -0.308383449
[21] -0.324212444 -0.337685552 -0.455131621 -0.625250219 -0.774801239
```

Figure 8

In the case of scree plot, we can identify the number of factors through the degree of drop or slope between the number of factors. According to the scree plot, there is large drop until the 4th factor but after that, the slope is very mild. (Figure 9) It means the scree plot suggests including 4 factors.

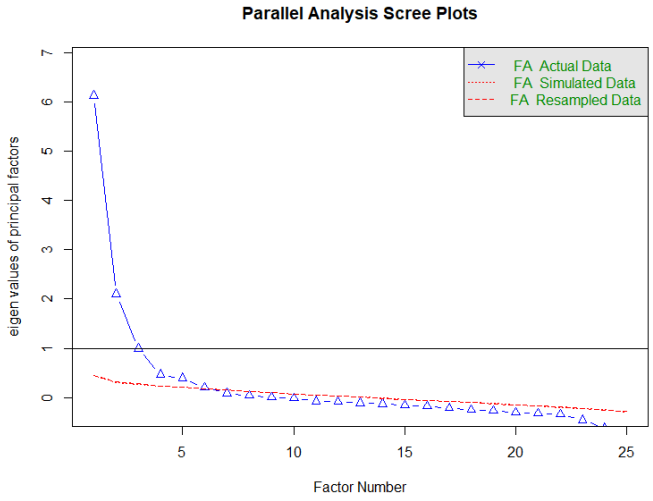


Figure 9

The model below is the result of factor analysis with 3 factors included without rotation. (Figure 10) ML1 indicates the reason for sport is for physical appearance. ML2 is about the competitive spirit during the sports. ML3 affects to questions about learning new skills and developing them. However, it was hard to find connection for several questions to its assigned factors.

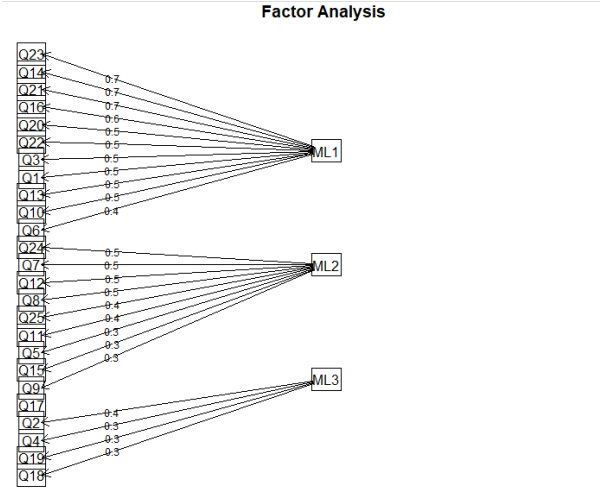


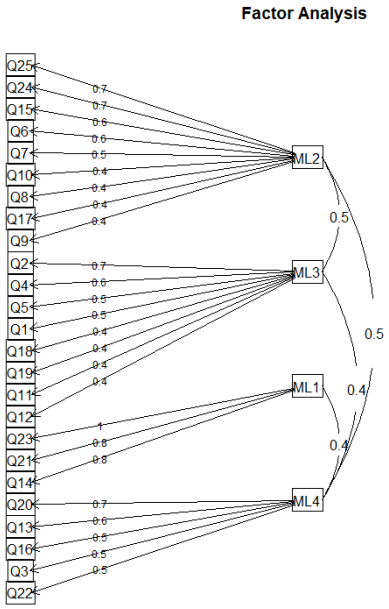
Figure 10

Simple structure

Below diagram with the value of loadings are from the EFA model with 4 factors, maximum likelihood as fitting estimation and oblimin for rotation method.(Figure 11) Since the previous model with 3 factors has difficulties in interpretation with a few variables, I added one more factor as scree plot suggested. The model with 4 factors is providing more precise interpretation to explain the reason of sports without much less vagueness compared to the model with 3 factors. Followings are 4 main reason why people enjoy sports according to factor analysis:

- ML1: Physical health improvement
- ML2: Joyfulness after winning competition or others
- ML3: New skills acquired with self-motivated practice
- ML4: Better appearance

Even though Q9, Q11, Q12, and Q17 have loadings lower than cutoff point which is 0.4, all of those 4 questions were aligned very well to explanation of factors. Hence, there were no bad questions to be removed.




```
Loadings:
    ML2    ML3    ML1    ML4
Q1          0.522
Q2          0.708
Q3                        0.515
Q4          0.577
Q5          0.526
Q6    0.603
Q7    0.538
Q8    0.418
Q9
Q10   0.434
Q11
Q12
Q13                       0.644
Q14                0.771
Q15   0.616
Q16                       0.539
Q17
Q18         0.446
Q19         0.419
Q20                       0.705
Q21                0.805
Q22                       0.512
Q23                0.977
Q24   0.656
Q25   0.702

                 ML2   ML3   ML1   ML4
SS loadings    2.865 2.333 2.321 1.998
Proportion Var 0.115 0.093 0.093 0.080
Cumulative Var 0.115 0.208 0.301 0.381
```

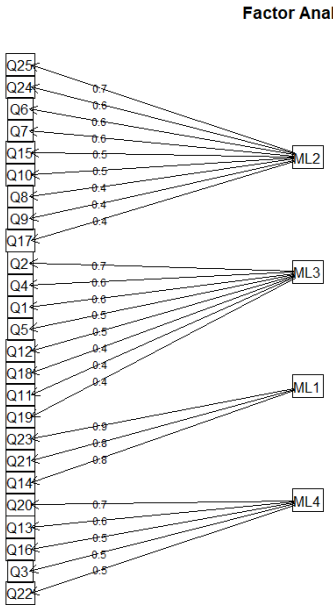
Figure 11

In addition to the previous model with oblimin as rotation method, I built another model which includes 4 factors, maximum likelihood as fitting estimation and varimax for rotation method. (Figure 12). The only difference I could find between model using oblimin and model using varimax was this new model is not allowing correlations between factors while oblimin does. Although loadings for each question were slightly different to the previous model, there was no

change in interpretation because exactly same questions were assigned to exactly same factors compared to the previous model. The new model also shows 4 factors to explain the reason of sports as follow:

- ML1: Physical health improvement
- ML2: Joyfulness after winning competition or others
- ML3: New skills acquired with self-motivated practice
- ML4: Better appearance

When it comes to loadings, Q19 was the only question that carries the loading lower than 0.4. However, Q19 still helps to interpret that some people play sports because they want to earn new skills and develop them thorough self-practice. Hence, Q19 will not be removed.




```
Loadings:
    ML2    ML3    ML1    ML4
Q1          0.555
Q2          0.682
Q3                        0.529
Q4          0.562
Q5          0.517
Q6   0.578
Q7   0.567
Q8   0.438  0.402
Q9   0.410
Q10  0.492
Q11         0.438
Q12  0.417  0.498
Q13                       0.603
Q14                0.775
Q15  0.522
Q16                       0.543
Q17  0.401
Q18         0.449
Q19
Q20                       0.660
Q21                0.798
Q22                       0.513
Q23                0.946
Q24  0.645
Q25  0.651

                 ML2   ML3   ML1   ML4
SS loadings    3.113 2.898 2.413 2.288
Proportion Var 0.125 0.116 0.097 0.092
Cumulative Var 0.125 0.240 0.337 0.428
```

Figure 12

Write up

a. The survey introduced the Likert scale to analyze the reason why people play sports. The scale is from -3 to 3 with 0 as neutral reply.

-3	-2	-1	0	+1	+2	+3
Stronly disaree	Disagree	Slightly disagree	Neither agree nor disagree	Slightly agree	Agree	Strongly agree

b. Factor analysis was used to interpret the survey result with several methods during model building
    - Rotation: none, oblimin, varimax
    - Fitting estimation: maximum likelihood (other possible options: minres, uls, wls, gls, pa)
    - Program used: factor analysis (other possible options: pc, both)

c. Following is the number of factors suggested by each method:
    - Parallel analysis: 6
    - Old Kaiser criterion: 2
    - New Kaiser criterion: 3
    - Scree plot: 4

    I decided to use 4 factors because after trials with 2 and 3 factors, it was unable to result in good interpretation that makes sense. 2 and 3 factors were too less to identify characteristics of each factors clearly. On the other hands, 6 factors provide too much detailed segmentation of questions, which have it hard to interpret the survey replies.

d. I tried two fa models, both of which use maximum likelihood and have 4 factors. Only difference is that one is using oblimin rotation and the other is using varimax rotation. Coincidently, there were no questions that need to be removed since, even with loading values smaller than cutoff, all questions contributed to find out unique characteristics of each factors.

e. Final model is the model which includes 4 factors, maximum likelihood as fitting estimation and varimax for rotation method. (Figure 12)

f. Interpretation of the factors:
    - ML1: Physical health improvement
    - ML2: Joyfulness after winning competition or others
    - ML3: New skills acquired with self-motivated practice
    - ML4: Better appearance

g. Adequacy of solutions

    - Fit indices: CFI (Comparative Fit Index)
      Since a model with CFI > 0.90 is acceptable and CFI > 0.95 is good, the final model with 0.93 of CFI is acceptable. (Figure 13)

```
> 1 - ((finalmodel$STATISTIC-finalmodel$dof)/
+        (finalmodel$null.chisq-finalmodel$null.dof))
[1] 0.9303064
```

Figure 13

- Reliabilities: Cronbach's alpha
  Cronbach's alpha provides an estimate of how much items are hung together. Since the value of 0.70 or 0.80 of Cronbach's alpha is acceptable, the final model is also acceptable.

```
Reliability analysis
Call: alpha(x = nooutlier[, factor1])

  raw_alpha std.alpha G6(smc) average_r S/N    ase mean  sd median_r
      0.88      0.89    0.86      0.74 8.4 0.0073   -2 1.3     0.77

 lower alpha upper     95% confidence boundaries
0.87 0.88 0.9


Reliability analysis
Call: alpha(x = nooutlier[, factor2])

  raw_alpha std.alpha G6(smc) average_r S/N   ase mean   sd median_r
      0.79      0.83    0.82      0.35 4.8 0.011  2.1 0.76     0.34

 lower alpha upper     95% confidence boundaries
0.77 0.79 0.82


Reliability analysis
Call: alpha(x = nooutlier[, factor3])

  raw_alpha std.alpha G6(smc) average_r S/N   ase mean   sd median_r
      0.76      0.78    0.78      0.31 3.6 0.013  2.5 0.53     0.27

 lower alpha upper     95% confidence boundaries
0.73 0.76 0.78


Reliability analysis
Call: alpha(x = nooutlier[, factor4])

  raw_alpha std.alpha G6(smc) average_r S/N   ase mean  sd median_r
      0.78      0.78    0.75      0.42 3.6 0.013 0.65 1.2     0.42

 lower alpha upper     95% confidence boundaries
0.75 0.78 0.81
```

Figure 14