Discussion of Results

Question C

Based on the eigenvalues >= 1, we extracted 7 factors. However, after testing 5, 6, and 7 factors using oblimin (an oblique rotation method) and varimax (an orthogonal rotation method), the 5-factor (in both method cases) had the lowest number of cases of cross-loadings and low loadings.

Question E

Based on the silhouette score, the K-Means ($R2 = 0.509$) performed better than the Gaussian mixture model clustering technique ($R2 = 0.307$). A look at the scatterplots (scatterplot_kmeans_5_png and scatterplot_gmms_5_png) generated from the first two columns of the dataset for both techniques shows that the K-Means produced more distinct and well-separated clusters.

Question H

To select the 20 questions, I ran a correlation analysis between the math ability measure and each of the 40 questions. I then sorted the correlations by absolute value and printed the top 20 questions, as shown below.

To determine the 30-individual team, first, if there are reverse-coded questions among the 20 questions, I would subtract the corresponding responses for each participant from 6. For example, if a participant selected 1 for reverse-coded questions, I would change the response to 5. After this step, I would sum up the responses to the selected 20 questions. The top 30 participants will be selected to form the team.

Question 1: Q17
Question 2: Q15
Question 3: Q36
Question 4: Q4
Question 5: Q2
Question 6: Q3
Question 7: Q11
Question 8: Q10
Question 9: Q14
Question 10: Q18
Question 11: Q24
Question 12: Q28
Question 13: Q32
Question 14: Q6
Question 15: Q26
Question 16: Q34
Question 17: Q39
Question 18: Q22
Question 19: Q12
Question 20: Q38

Question I

For the other project, there are better choices than using the above questions (obtained from correlation with math ability). Since the selection will be based on a variety of many different personality traits, a good solution is to average the absolute factor score for each of the 40 questions. The top 20 questions based on the average are selected (see below). It can be seen that some of the questions below are not among those for the math project (above).

Question 1: Q17
Question 2: Q15
Question 3: Q5
Question 4: Q13
Question 5: Q34
Question 6: Q22
Question 7: Q36
Question 8: Q39
Question 9: Q35
Question 10: Q20
Question 11: Q32
Question 12: Q8
Question 13: Q27
Question 14: Q4
Question 15: Q16
Question 16: Q11
Question 17: Q31
Question 18: Q2
Question 19: Q9
Question 20: Q14

Bonus: Question M

Yes, it is possible to use the questionnaire answers to predict the individuals' math ability. One technique that can be used to do this is multiple regression. However, it is not a good idea because the number of predictors that can be used effectively in a multiple regression model depends on the complexity of the relationship between the predictors and the response, as well as the size of the dataset.

Using a large number of predictors in a predictive model can increase the risk of overfitting, especially if there is a high degree of collinearity among the predictors. Collinearity among predictors refers to the situation where two or more predictors are highly correlated with each other. Considering that the dataset relates to personality traits, it can be expected that some of the variables will be highly correlated with other variables.