1. You are given a dataset with 21644 individuals. The dataset contains the answers to a questionnaire with 40 questions to evaluate their personality traits and a measure of the math ability of the individuals. Your task is to cluster these individuals into groups and relate the personality traits to their math ability.

d/e) Which algorithm gives you a better result? Explain your answer or explain why it is not possible to evaluate which algorithm is better.

I think Gaussian gave me the better result. As it is based on the normality concept and separate the groups into four as shown in the figure below. I do not think we can validate the Gaussian or k-mean.

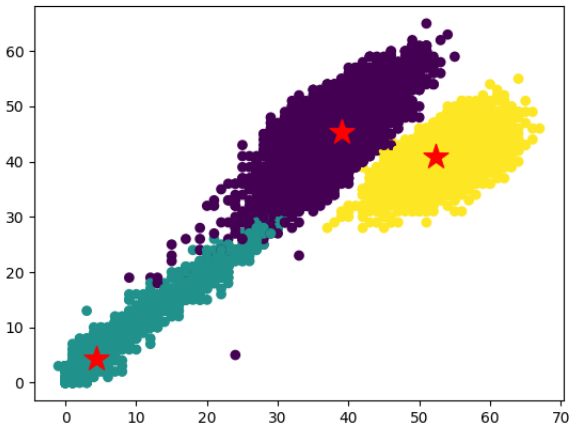

Figure1: Representation of different group using Gaussian method.

However, the k-mean gave the other results,

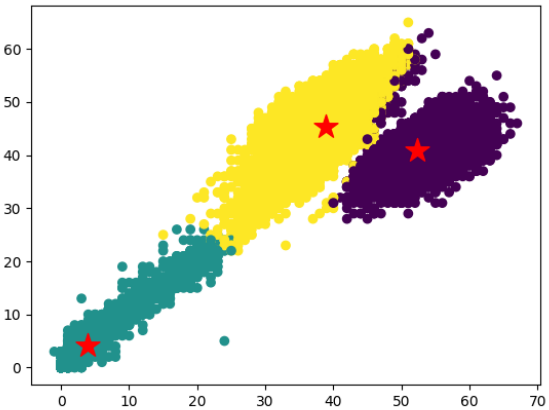

Figure2: Representation of different group using K-mean.

f/g) Use the personality traits to predict the math ability of the individuals. You may use linear regression, logistic regression, or any other supervised learning techniques.

Ans = I did linear regression after factorial analysis. I put math as a dependent variable and other factor as an independent variable and then I analyzed the data. I found the R2 of pretty low around 0.12 in case of linear regression. I don't think we can use logistic model to predict math it is because, Logistic is for the categorical dependent variable. Because of that I generated the dummy variable for the math score. I put 1 for math score greater than 100 and 0 for score less than 100. After this I got the accuracy score of 0.78, however before converting to dummy I run the logistic regression and got the score of 0.04, which is pretty low.

h) Now you are assembling a team of 30 individuals to work on a math project. You want to choose the individuals with the best math ability. However, you cannot choose those people who are in the original dataset. You can only choose 30 individuals from the population. Also, you do not have the resources to do a math test nor to collect 40 answers from those new recruits. You can only collect 20 from them. Which 20 questions should you choose among the 40 questions in the original questionnaire? And how will you use the information you collect from this new questionnaire to assemble your team? Explain your answer.

Ans = The 20 questions with maximum variance are 'Q26', 'Q10', 'Q19', 'Q16', 'Q15', 'Q12', 'Q34', 'Q3', 'Q18', 'Q24', 'Q23', 'Q39', 'Q11', 'Q9', 'Q17', 'Q22', 'Q28', 'Q7', 'Q5', 'Q8'. These questions are likely to cover broad spectrum of math skills and could be most indicative of math availability. I will collect the information based upon these questions, after I am done with collecting, I will then do the variance of the new data. The least variance performs the best and I will sort the variance and select the top 30 candidates.

i) Suppose instead of a math project, you are assembling a team of 30 individuals to work on a project that requires a variety of different personality traits. Which 20 questions should you choose among the 40 questions in the original questionnaire? Is your answer different from the previous question? Explain your answer.

Ans = For a diverse set of personality traits, I would want to identify questions that cover a wide range of characteristics. The methods for selecting questions would involve choosing those that offer the most comprehensive coverage across various personality dimensions. Instead of solely looking for high variance, I consider using techniques like factor analysis or looking for questions that are well established in different personality traits. Hence the selection process for personality-based questions would differ significantly from selecting questions for assessing math abilities. The focus would be on capturing a bread array of personality traits rather than variance in specific skill sets. The questions I would select for this problem are: 'Q34', 'Q39', 'Q17', 'Q15', 'Q16', 'Q26', 'Q8', 'Q4', 'Q5', 'Q10', 'Q24', 'Q35', 'Q36', 'Q13', 'Q40', 'Q22', 'Q27', 'Q20', 'Q18', 'Q9'. These are the questions that cover a broad spectrum of personality traits based on their contribution to the extracted factors from the questionnaire data. Adjusting the number of factors can refine the selection based on the granularity of personality dimensions.

Bonus question: It is also possible to use the questionnaire answers themselves to predict the math ability of the individuals. Explain whether this is a good idea or not and why.

Ans = I don’t think it’s a right thing it is because some might exaggerate and some might just mislead their capacity which might cause bias in analysis.