ECON860 Data Analysis For Economics(Fall 2023)
Final Exam 1
Due: 17th December 2023 (1:00pm)

Name: Grant Hicks
UID: C10857891

---

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.

   (a) The questionnaire is similar to the "Big Five Inventory" in the lecture (But not the same, so they do NOT necessarily correspond to the Five personality traits we mentioned in the lecture). However, you do not have the cookbook, so you do not know which questions correspond to which personality traits.

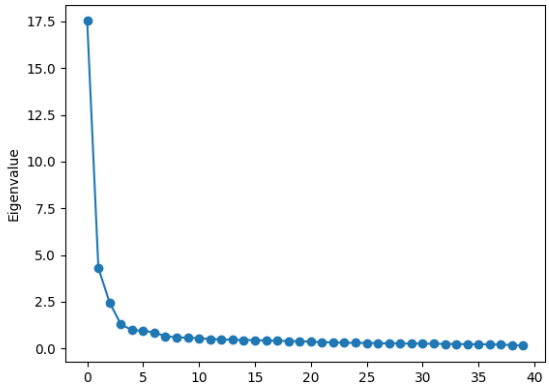
   (b) The answers are on a scale of 1 to 5. If the individual refuses to answer that particular question, the value would be 0.

   I chose to include the question values of 0 in my dataset (therefore there are 6 answers, scaling from 0 to 5). I could have written my program to filter them out, but my thought was that individuals who chose not to answer part or all of the questionnaire could be just as insightful as the answers themselves, especially because individuals with all 0 responses seemed to be disproportionately skewed to those who had exceptionally high or exceptionally low math scores. The data confirms this – I thought these contrarians were noteworthy for clustering different personalities.

   (c) Use factor analysis to get a measure of several personality traits from the questionnaire. Notice that this questionnaire is not the same as the "Big Five Inventory", so you may not have exactly five traits. You need to follow the procedure introduced in the lecture to find out what is the suitable number of traits (factors).

   Based on the eigenvalues returned from the unaltered dataset, I chose to retain 2 factors (or measure 2 personality traits). It would be reasonable to make an argument for retaining 3 or 4 factors, but based on the eigenvalues, the first factor explains 43.81% of the variance. A second factor brings it to a cumulative 54.58% of the variance. Adding a third and fourth factor only explains a cumulative 60.75% & 64.01% of the variance, respectively. Based on these factor statistics, I didn't consider adding a third or fourth factor to be worth the noise they introduce to the data. With more context beyond raw numerical data, my choice may have been different.

   Suppressing the 0's in the initial dataset would certainly change this approach, as the 0's are largely responsible for the substantial first eigenvalue. If the 0's were suppressed, 5 or more factors could certainly be appropriate based on eigenvalues.



(d) Use the personality traits to cluster the individuals. You may use KMean clustering, Gaussian mixture model, or any other unsupervised learning techniques.

I tested KMeans, KMedoids, and Gaussian mixture model

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

Gaussian mixture model, with 3 clusters, is the best result, because it is better suited to non-spherical clusters (these clusters are more ellipse-shaped). This is especially practical in differentiating the individuals within the dataset whose low trait scores imply they did not complete some, or all, of the questionnaire. Despite the 2 cluster GMM having the highest silhouette score, I would elect to use the 3 cluster GMM. This is because while the program may calculate that the clusters are more distinguished when the top right quadrant of the scatterplot is one cluster, human eye sees there are two distinguished ellipses in that quadrant.

(f) 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.

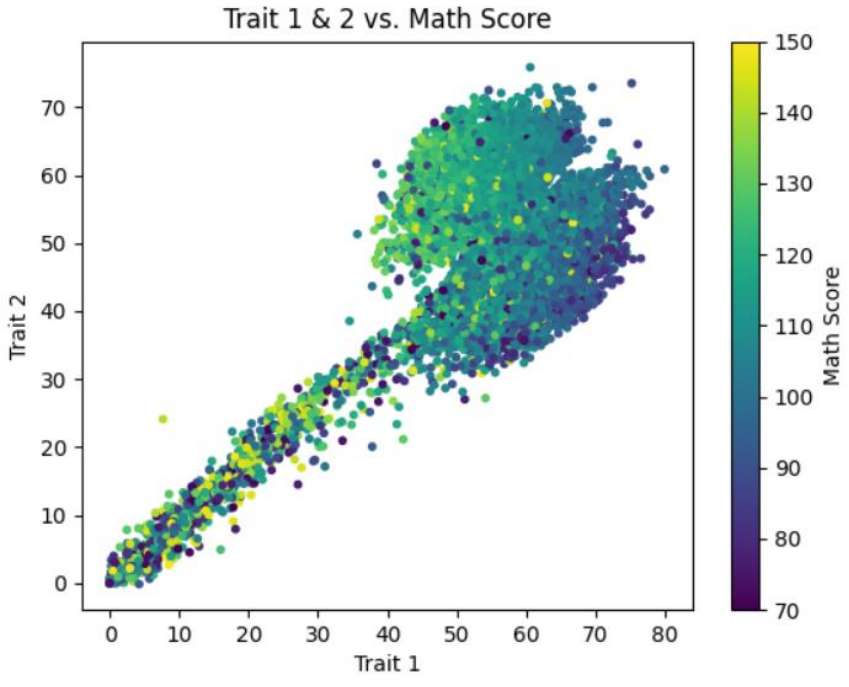
I tested linear regression & logistic regression.

(g) Which model gives you a better result? Explain your answer or explain why it is not possible to evaluate which algorithm is better.

The linear regression model gives a better result since the traits produced from factor analysis are continuous variables. That said, the R-Squared score is only approximately 0.12, so the fit of the model may be weak, though it is hard to say for certain if it's a good score without more context.

I included the logistic model to show it was attempted, but it is not usable. The R-Squared score is negative, meaning the prediction is worse than if the simple mean of the training data was used.

(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.

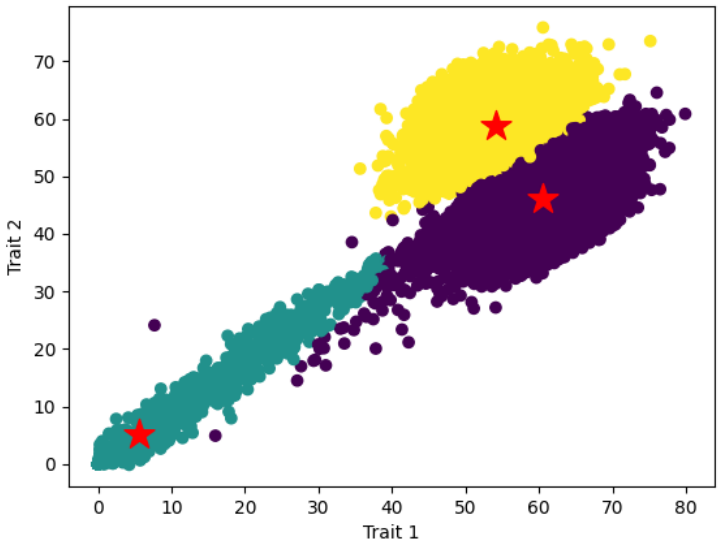


The above plot tells me that, ideally, I want to seek individuals with value that are both approx. [40-55] in trait 1 & approx. [45-70] in trait 2, as that is where I can find the greatest concentration of exceptionally high math scores.

To find these individuals from the population, I will use the factor loadings I saved into a csv in 'run_factor_analysis.py'. This data tells me which questions have the largest effect on trait 1 (column title 0) & trait 2 (column title 1). I converted this dataset to absolute values since values closest to 1 or -1 are most impactful, to raise or lower trait scores, respectively. The 10 questions with the highest loading score for trait 1 and the 10 questions with the highest loading score for trait 2 will be used. With these 20 questions, I can be confident that my assessment of each individual for trait 1 & trait 2 is reasonably accurate. Then I can target the range mentioned above based on the original dataset.

Rank	Factor 1		Factor 2	
1	Q17	0.87813	Q34	0.844937
2	Q15	0.848013	Q39	0.835109
3	Q4	0.711826	Q8	0.801237
4	Q10	0.703076	Q22	0.794094
5	Q11	0.676597	Q16	0.765148
6	Q36	0.673022	Q13	0.754866
7	Q2	0.671766	Q5	0.748556
8	Q3	0.663188	Q20	0.744777
9	Q18	0.649394	Q27	0.713252
10	Q38	0.644184	Q9	0.609134

(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.



Based on the 3-cluster Gaussian mixture model, if I wanted to assemble a team of 30 individuals with a variety of personalities, I would want to focus all 20 questions on being able to evaluate trait 2. If I can evaluate trait 2 accurately, then based off the trait 2 values, I should be able to take the top 10, median 10, and bottom 10 values and get several individuals from each of the 3 clusters, if not exactly 10 from each.

Conversely, I could choose to take the top 5 (yellow cluster), median 5 (purple cluster), and randomly select 20 individuals from the bottom quartile of trait 2 values (blue). The main commonality of the blue cluster is partial or total incompleteness of the questionnaire. Since we can see a clear connection between math score and traits in the other clusters (referenced in (h)),

and there is incredibly high variance in the math scores of the blue cluster, we are likely to have very high variance in the personalities of those in the blue cluster as well. For that reason, it may be beneficial to weight our selection more to that cluster for the best chance at maximum personality diversity, while still ensuring we get a few individuals from each of the other two clusters.

Top 20 factor loadings to evaluate trait 2:

Rank	Factor 2	
1	Q34	0.844937
2	Q39	0.835109
3	Q8	0.801237
4	Q22	0.794094
5	Q16	0.765148
6	Q13	0.754866
7	Q5	0.748556
8	Q20	0.744777
9	Q27	0.713252
10	Q9	0.609134
11	Q26	0.597413
12	Q31	0.592248
13	Q12	0.578323
14	Q24	0.542419
15	Q32	0.504751
16	Q19	0.468649
17	Q40	0.445838
18	Q30	0.444181
19	Q28	0.444001
20	Q23	0.439888

(j) You must hand in your homework via Github. Create a repository named "ECON860_final". In your repository, you should have the code and a .gitignore file. You should also include a file named README, which includes step-by-step instructions on how to run your Python code to collect the data you collected. This is especially important if you have multiple Python files. You should have a written answer committed in your repository. It can be included in the README file or it can be a separate file.

(k) You can commit and push to Github as many times as you like. Only your last commit before the deadline is graded. I can read your previous commits, but they will not be graded.

(l) It is more important to hand in partial work than to not hand in anything. For example, if you are not able to get a nice set of personality traits in earlier parts, you can still hand in the code you used to cluster the individuals. Also, you can use whatever imperfect traits you have to predict the math ability of the individuals. You will get partial credit for the parts you have done.

(m) 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.

While it is possible, the confidence level of the prediction depends on whether all 40 question answers for the individual whose math ability we are attempting to predict is at our disposal. This is because our 'machine.loadings_' in 'factor_analysis.py' mostly show low magnitude on trait scores on a stand-alone basis (other than 0/incomplete – but we can't predict math score for these individuals because those with incomplete questionnaires had widely varying math scores, as previously mentioned). With a completed 40 question response though, we can be more confident in the cumulative significance of the machine loadings and more confident in our prediction for those who completed the questionnaire with no omissions.