ECON 8600 Midterm
Elizabeth Zarrilli

** Note: For a more detailed description of how to run the individual python files to answer the question, see the github repository .README. This report will focus mainly on reporting the results and answering the questions.

This project asked to use factor analysis to reduce a list of individuals' responses to 40 questions on a personality test into an unspecified number of personality traits. Then use clustering techniques to group the individuals based on these traits and predict math scores using supervised learning techniques.

Part 1: Factor Analysis

The question asks to use factor analysis to condense numerical responses (1-5) to 40 personality questions into personality traits. Because we have an unknown number of traits that the test identifies, we first need to determine this correct number. Then we want to condense the information from individuals' responses to the 40 questions into variables for the identified traits.

a. Before starting I removed observations with missing questionnaire data (any response = 0). I first did this for only those who did not answer any of the questions but ultimately decided that any refusal to answer the questions may bias the results so I ultimately chose to remove any individual who did not answer all of the questions. This reduced the dataset down to 15195 observations. I called this DataFrame 'df_no_zeros'.
    i. I did actually try to do factor analysis before removing the zeros and only ended with four traits with eigenvalues > 1. This code is commented out but the incorrect results dataset is saved as results4.csv.
b. I separated out the math scores from the personality question responses to only work with the data to determine the personalities.
c. I checked to make sure that the variance matrix of the 40 questions was statistically different from an identity matrix using the Bartlett sphericity test. The p value was about 0.0.
d. I generated the eigenvalues for the data and the first seven were > 1. This indicates that the correct number of factors is seven.
e. I calculated factor loadings using both no rotation, which did not give very distinct correlations and 'varimax' rotation which gave much more distinct correlations (each question was much more correlated with one factor than the others). These are the loadings I used to generate my result data. Using 'varimax' or non-orthogonal rotation was also motivated by the fact that personality traits are likely correlated with each other in some way and so the resulting factors should not be constrained to be orthogonal to each other.

f. I then took the dot product of the questionnaire data and the factor loadings and saved the result in a csv file called ‘results_no_zeros.csv’.

Part 2: Grouping the Respondents

Here we were to group the observations into clusters using an unsupervised learning model.

a. Use the dataset from part 1 with individuals’ personality traits, but removing their math scores since this is an unsupervised model, to cluster the individuals.
b. I ran programs for both gaussian and KMeans clustering models to group the individuals by personality traits. For both methods, I ran a function with n = 2,3,...10 groups and generated their silhouette scores. I plotted these on a pyplot and found that for both methods, n=4 maximized the silhouette score at around .5. After running both I chose to use KMeans clustering because the two methods did not seem to differ much in their results and there did not seem to be any specific reason to use Gaussian Mixture.
c. I grouped the individuals into four clusters using this technique and coded their cluster IDs as a variable called ‘cluster’ in the dataframe and saved this dataset as ‘clustered_data.csv’.

Part 3: Supervised Learning Model to Predict Math Scores

a. Since math is not a binary or categorical variable I did not feel logistic regression would be the most appropriate choice of supervised learning model for this problem. Therefore, I used a linear regression model to predict the individuals’ math scores. Although I did run a logistic k_fold template to test it and got much lower R2 scores. However, when I decided not to use these results I deleted the code for the logistic regression.
b. First I scrambled my data before splitting it into training and test data.
c. I used a k_fold template to run the linear regression with 4 splits. My target was math scores and my data were the personality traits from part 1. This template splits the data into training (75%) and test (25%) observations. Each round returned an R2 of about .97/.98 which indicated the model explains a good amount of the variation in individuals’ math scores and had relatively good predictive capacity.
d. I also ran the regression using the dummy variables of the cluster groupings from part 2 though this returned much lower R2 values (~.6) so it was not as predictive of a model.
e. Using a linear model without any splits, I generated predicted math scores for all individuals in my sample. These predictions appeared to be pretty accurate to the true scores. These values were added to the dataset as a variable called ‘predicted_score’.
f. Linear regression is a more appropriate model for this data because logistic regression is primarily used for binary or categorical data. Because the math scores are numerical data, I used linear regression.

Part 4: Selecting 20 Questions to Target Individuals with the Highest Math Scores

a. The goal of this question is to identify individuals by personality traits that are most likely to be good at math. I used the seven traits I identified by factor analysis to see how they correlate/predict an individual's math score.
b. To do this, I generated the coefficients for the linear regression of the seven traits on math score. This returned a coefficient matrix:

Trait	coefficient
x1	1.323940
x2	0.113371
x3	-2.434674
x4	0.024061
x5	1.075481
x6	1.785166
x7	0.548143

It can easily be seen that traits 6, 1, and 5 have strong positive effects on math score while trait 3 has a strong negative effect on math score. This means I need to ask the 20 questions that will most identify traits 6, 1, 5 and 3.
c. Using the factor loadings from my factor analysis I pulled the questions with the strongest correlations to traits 3, 6, 1, and 5, since those have the greatest impact on an individuals' math ability.
    i. The questions I would select by this method are: questions 7, 32, 14, 24, 26, 12, 8, 16, 23, 22, 39, 3, 11, 15, 17, 23, 40, 13, 6, 1. Individuals' responses to these questions are the most correlated with the 4 most predictive traits for one's math score.
    ii. I will select individuals who answer highest on the questions positively correlated with traits 6, 1 and 5 and lowest on the questions negatively correlated with trait 3. As well as those that answer lowest for questions with the converse correlation.

Part 5: Identify Individuals with Diverse Personalities:

a. Since this project does not require strong math ability, we do not care at all about how the selected questions predict one's ability to do math. Instead, I will look for the questions that most strongly identify all seven of the personality traits. The only difference between this method and the previous is that I am not weighting the traits that strongly predict math ability over others. The goal will be to select questions that can strongly identify all 7 traits.
b. Similar to the previous I looked at the factor_loadings of the seven traits and found the twenty with the strongest correlations, however this time I looked at the correlations for all traits, rather than restricting it to just those that predict math ability.

c. The resulting list of questions is: 1, 5, 7, 8, 9, 12, 13, 15, 16, 21, 22, 23, 26, 27, 31, 34, 38, 39 and 40.
d. You can then select individuals that rate themselves highly for the questions included above that are highly correlated with each of the 7 traits.
e. Additionally, here, you may want participants who fit very neatly into one of the different traits rather than having to satisfy all seven at once. In part 4 we wanted the chosen 30 to score highly on all the questions associated with traits 6, 5 and 1 and low on 3. But here there is no need for each individual to necessarily align with all traits, rather we want some individuals who fit strongly into each trait.

Part 6: Bonus Question

Though it is possible to use the questionnaire answers immediately to predict the individuals' math scores, I would not recommend this method. (Out of curiosity, I quickly ran this using a kfold_template and the R2 < .2 which was honestly even lower than I expected).

a. My instinct is that this model heavily suffers from multicollinearity. Since the questionnaire has multiple questions to help identify each targeted personality trait, several of the questions will explain the same variation in an individual's personality. Since the coefficients in linear regression models only explain the unique variation in y explained by each variable, if many of the questions identify the same personality trait, say extroversion, then there is not much unique variation to model. If questions 1 and 5 both are intended to tease out extroversion then we would expect an extrovert to score highly on both and an introvert to score low on both. This is similar to including both height in inches and height in feet in a linear model. Though that is an example of perfect multicollinearity (the model would not even run in that case), each question is likely highly correlated with at least one other question, introducing multicollinearity into the model.
b. Here we are testing if *personality traits/personality* impacts or is predictive of math ability. The questionnaire is only meant to assess these traits. The reason we might not just ask people directly "are you extroverted?" etc. is because they may not, even if unintentionally, give a good description of their personality. Instead the questionnaire asks situational questions to determine the traits and in this case the more questions connected to each trait the better to determine patterns. The test likely relies on correlation between responses to determine the respondents' personality. The questionnaire is probably designed in the opposite way as a linear model. This is why using factor analysis to group the information from the questionnaire answers is very important for linear regression and prediction in this problem.
c. Additionally, unrelated to the design of the questions themselves, but personality traits are likely not independent of each other. Being extroverted is likely correlated with how empathetic or patient etc. someone is. This is why it was better to use a non-orthogonal rotation in the factor analysis. We don't expect/require each trait to be uncorrelated with

the others. So even if each question identified a perfectly unique aspect of the respondent's personality, the answers are still likely to be correlated just because of the nature of personality traits. Factor analysis allows us to take this into account when reducing the variables.