This data task asks you to do 4 things using raw data from an experiment Josh conducted where children were given scholarships before kindergarten (many variables have been removed and some values changed to preserve anonymity). There are three CSV files:

1. baseline.csv which includes the children’s ID number, their baseline test score and their treatment status (whether they were offered a scholarship).
2. endline\_parent.csv which includes the children’s ID numbers and a variable with years of education that the main guardians have obtained as measured two years after baseline.
3. endline\_child.csv which includes the children’s ID numbers and a series of indicators for whether the children got one of the questions on a Matrix Reasoning test conducted two years after baseline correct.

You may use any software you like to do the following four tasks. When you are finished, please supply your commented code and a write up of your answers.

1. Please provide a table describing the attrition between baseline and endline
2. Please calculate the percentage of guardians who have less than 8 years of education
3. Please graph the percentage correct on each of the matrix reasoning questions to see whether the questions become more difficult as the test progresses (this is what is supposed to happen on the test).
4. Please generate the average percent correct on the matrix reasoning test, regress this on treatment controlling for baseline test scores, and interpret the regression in a few sentences.