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

1. childdata.csv which includes the children’s ID number, age, their test scores, their treatment status (whether they were offered a scholarship), and whether the data was collected at baseline or the endline which occurred after the children had attended preschool for two years.
2. guardiandata.csv which includes the children’s ID numbers, an indicator for endline, an indicator for whether the observation is for the respondent or the child’s other guardian, guardian’s education level, guardian’s preference for type of primary school, and the hours the guardian typically works.

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 baseline balance.
2. Please calculate the percentage of guardians who disagree about the type of primary school they are planning for their children
3. Please graph the distributions of the total score variable at baseline and endline by treatment status and interpret the graph.
4. Please group the subtests thematically and create z-score summary indices for each group following the procedure in Kling Leibman and Katz (2007). Regress the endline summary indices on treatment while controlling for baseline age, mother’s education, and the baseline value of the index. Present these regression results in a table and interpret the results in a few sentences.