Project 2

Pisa Plus 2012-2013 Data

For this project we are going to be working with a subset of the the 2012-2013 Pisa Plus Study data. This data includes demographic variables as well as achievement indicators for mathematics and science, and numerous non-cognitive variables such as attitudes, self-concept, cognitive activation, work ethic, etc., for each student.

Exercises

1. Read the pisa1 and pisa2 files separately into R.
2. Assume the demographic data was collected at one time point, and the questionnaire/achievement data at another, therefore resulting in two data files. Some students from whom we collected demographic data did not participate in the second questionnaire, meaning there are some students in the first data set who do not have any data in the second dataset. We only want to include those students from data set 1 for which there are matching values in dataset 2.

- Join the two datasets so that you keep all rows from pisa1 where there are matching values in pisa2, and all columns from pisa 1 AND 2. (*Session 6*)

3. We only have the individual items for the "interest" construct (int_a_t1, int_b_t1, etc). Create a new variable called "interest" that is the overall mean score of all the interest items. (*Session 5, last semester*)

- Optional: How could we calculate the overall and item reliabilities for the interest items in R?

4. Calculate a correlation matrix of all the continuous variables (excluding the individual interest items) and assign it to the object "corrmat" (*Session 7*)
5. Create a visualize representation of this correlation matrix using ggcorrplot. Label each correlation with its value (*Session 7*)

- Inspect the correlations. Which variables are most strongly correlated with each other?

6. Visualize the distributions of math achievement (ma_wle_t2) and science achievement (sci_wle_t2) with a histogram. What does the distribution look like?
7. Visualize the different distributions of math achievement (ma_wle_2), first by gender, and then by school type (Gymnasium, Realschule, Other) (*Session 7*). What could we interpret from these plots?
8. Research Question 1: Does mean math achievement (ma_wle_t2) significantly differ according to school type? How would we statistically evaluate this? (*Session 7, last semester*)
9. Check if the assumptions for the ANOVA are fulfilled (*Session 7, last semester*). Which assumptions do we need to check?
10. Interpret the results of your ANOVA. What does the ANOVA tell us? What does it not tell us? (Is there any way we can get more information?)
11. Let's now focus only on the Gymnasium schools. Create a subset of the data where school type = Gymnasium.

12. Using the Gymnasiums subset, try running a multiple regression model and checking the assumptions (try using the "check_model" function from the performance package. For example, you could investigate whether interest and self concept predict math achievement. Or if you think some other variables would be interesting, try that! Practicing visualizing the regressions (*Session 6*).

13. Try to think of some more insights you could gain from the data through different visualization techniques. For example, creating stacked or juxtaposed barplots, faceted graphs, creating new variables, etc. Make one plot we haven't yet looked at and briefly describe what information it displays.