Assignment 4

Data Visualisation

Olivér Heicz

2025-10-26

Task 1

In this task we are working with the nlschools dataset. The data set features information on 11-year old pupils in the Netherlands. It contains 7,681 observations of students from 215 primary schools in the Netherlands. Each observation represents a student and includes information about their class characteristics and test performance, especially in language tests.

Variables:

- **lang**: Language test score
- **IQ**: IQ score of the student
- **GS**: Class size: number of 11-year old students recorded in the class
- **COMB**: students taught in a multi-grade class or single-grade class (binary: 0/1)
- **SES**: Socioeconomic status of student's family
- **class**: class ID

The dataset examines how each variable, such as the class size, the type of class has an effect on the individual's preparation and actual performance on language tests. On the plot below we are visualizing this relation between the characteristics and the test results.

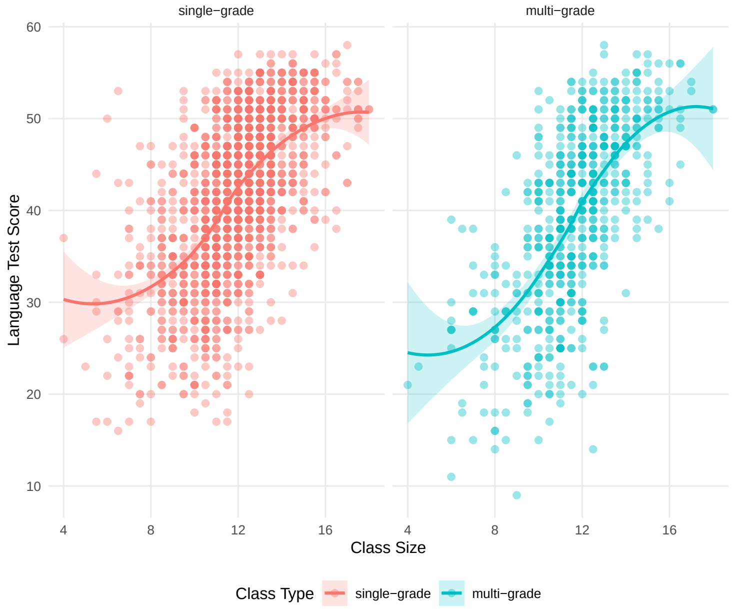

Figure 1: The relationship between class size, the type of class (single or multi-grade), and the score on the language test.

Looking at the plot, it reveals that greater language test points could be linked to a bigger class size, in contrary to what people would think in general. Multi-grade classes show higher improvement with bigger class size, maybe owing to the fact that students motivate each other creating a rich learning experience.

Task 2

In Task 1, we ignored the fact that the observations belong to different classes. This may be problematic because in each class the scores among students are related because they are studying from the same resources and taught by the same teacher and are influenced by each other. That is why, in this task we use the group-by-summarize method that creates unique groups which stands for each classes for every combination of class, GS, and COMB.

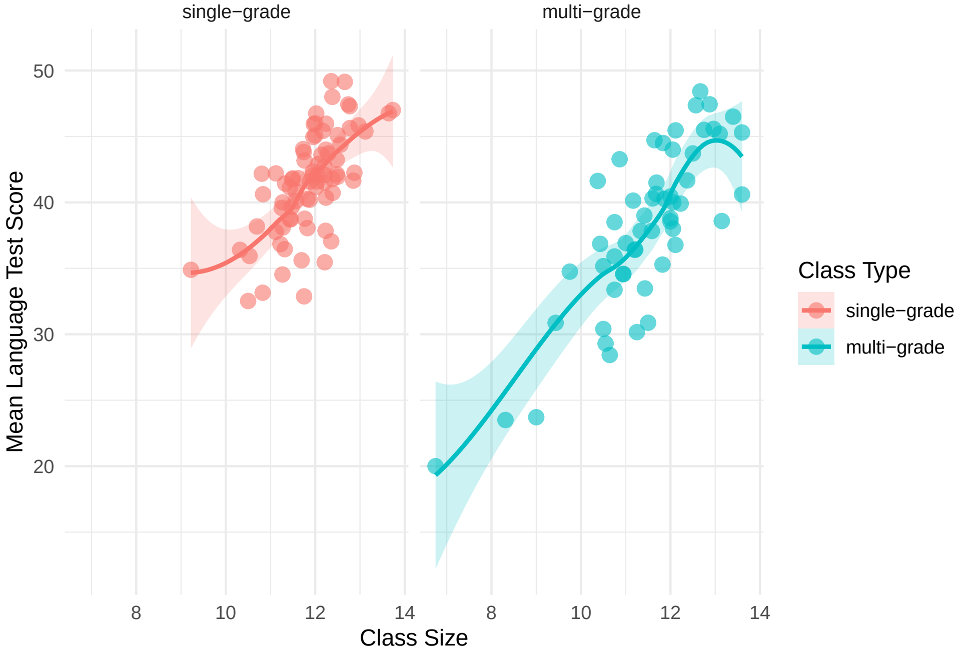

Figure 2: The mean of language scores for each class.

Task 1 plotted 7,681 individual students, with each point representing one student's language test score and their class's size. Now, the plot visualizes each class levels mean performance. We differentiate 215 classes. For both single and multi-grade classes a positive relationship can be seen. For single-grade classes this relationship is higher, classes have similar IQ ranges and as it increases, the performance on language tests shows high growth. Multi-grade classes show lower IQ ranges and a slower-moving improvement can be seen as well.

Task 3

In the final task we add the the SES - social economic status of student's family, and taking the same approach as in Task 1. For better interpretation we cut the new variable into three categories: Low, Medium, and High SES.

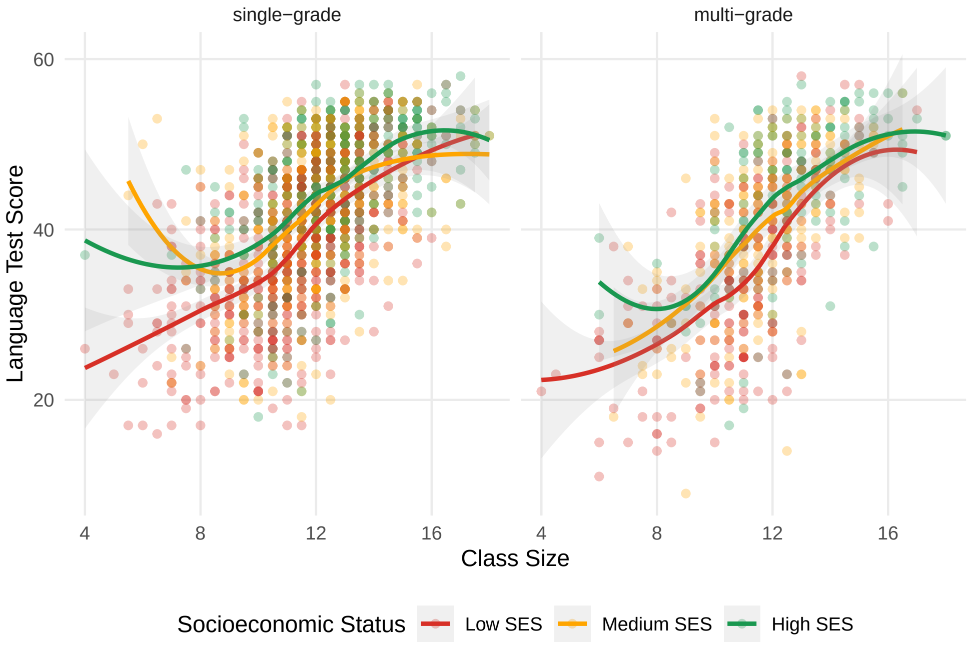

Figure 3: Language test scores based on the the class size, class type and the social-economic status of the student.

Adding SES creates quite different results compared to the other two plots. The social-economic variable has a big impact on the individual students performance. In both single-grade and multi-grade classes, high-SES students have better performance than low-SES and medium-SES students across all class sizes. This could be explained by those students can afford better resources such as private teachers, extra books and so on. The trend line of the highest SES is significantly higher than the others. Students from higher socioeconomic backgrounds maintain high test results independent of class size, indicating a flatter relationship between performance and class size. As for the low and medium SES students the trend line is steeper than high SES. It looks like that two groups benefit more noticeably from larger class sizes rather than the other SES statuses. In multi-grade classes, low-SES students show better improvement with class size, suggesting that mixed-age classrooms may provide some benefits for low socioeconomic status students.

Appendix

In this assignment I used AI only for clarification of R syntax questions, and for advice on formatting issues. All R code in this assignment was written by me.