Assignment 4

STAE04/DABN19: Data Visualization

Dirk Baars

2021-10-20

1 Task 1

This assignment contains the dataset 'nschools' from the MASS package and is collected by Snijders and Bosker in 1999 and used as a running example study of 2287 eighth-grade pupils (aged about 11) in 132 classes in 131 schools in the Netherlands. A table with the variables can be found here:

Table 1: Name, Type and Description of 'nlschools' dataset

Name	Type	Description
language.test.score	integer	Score of the language test
IQ	double	verbal IQ
class.ID	factor	ID of the class
class.size	integer	number of eighth-grade pupils recorded in the class
SES	integer	social-economic status of pupil's family
type.class	factor	were the pupils taught in a multi-grade class (0/1)?

Classes which contained pupils from grades 7 and 8 are coded 1, but only eighth-graders were tested.

Before examining the dataset, COMB is recoded, so that 0 is a single-grade class and 1 is a multi grade class. Then the variables lang, class, GS and COMB are recoded into language.test.score, class.id, class.size and type.class respectively to make the meaning of the variables more clear.

A scatterplot is created with geom_point, whereby the size of the class is plotted against the language test score. The plots are faceted against the type of class. To prevent overplotting, 'jittering' has been applied, and the points have been made more transparant with 'alpha'. To get better insight in the relation, a smooth line is added.

In figure 1 the scatterplot is found. There is a relationship visible between class size and test scores. The relationship is most visible in the multi-grade class, whereby the increasing of class size seems to go hand in hand with the increasing of the language test score. In the single grade class, the test score tend to be slightly decrease with an increasing class size. So there is an overall relationship and it differs between single and multi-grade classes. However, the grades are overall higher in a single class, so it seems to be most beneficial for students.

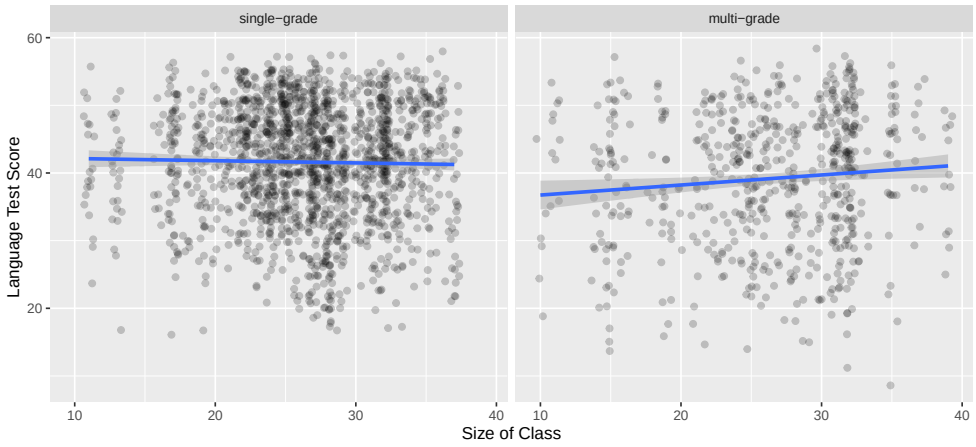

Figure 1: Association between language score test and the size of class, faceted by type of class

2 Task 2

In this task, the scores are summarized by computing the mean of the language tests, which are then plotted in the same matter as in the previous task. It seems like the relationship remains the same. On the downside, big amounts of data is lost, whereas in the first task there were 2287 observations and in the second there are only 133 left.

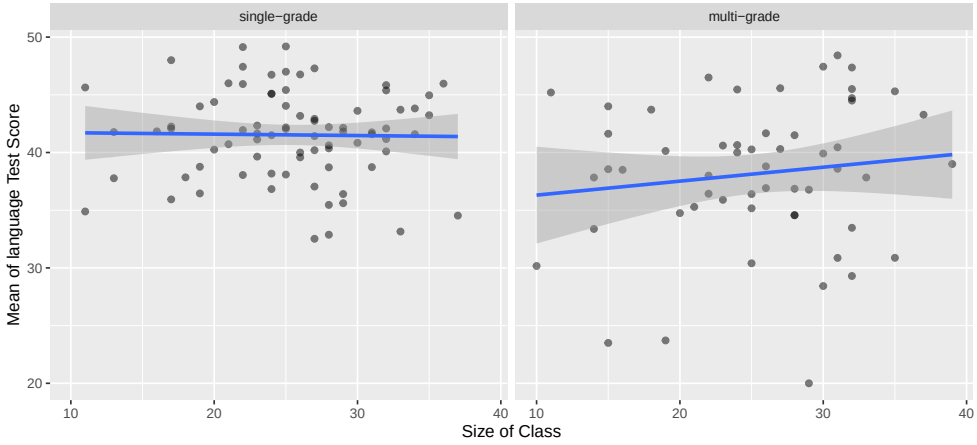

Figure 2: The mean of language test scores summarizes by class IDs across types of classes and different class sizes, a faceted scatterplot with added line.

3 Task 3

For the last task, the same scatterplot is created, however now there is also extra facets on social economic status. Whereby the SES variable is cut into an ordinal

variable, which exist of 'high', 'medium' and 'low'. . The plot is seen in figure 3 below:

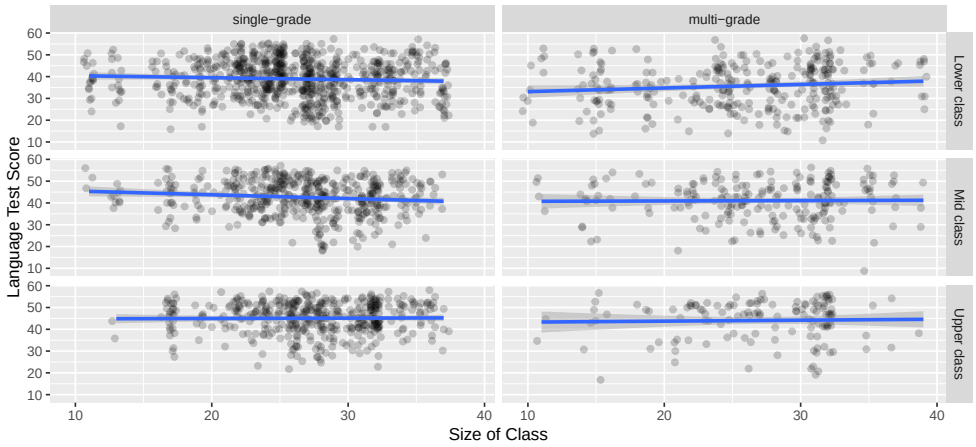

Figure 3: Language test scores across SES levels, class sizes and type of classes. A faceted scatterplot with added line.

Interesting to see, that overall in all single grade there is a small negative relation and in multi grade there is a small positive relation. In multi-grade, on average, the scores are higher for upper class, then mid class and the lowest for the lower class. In the single grade class, the lowest grades are with the lower class, whilst the mid and upper class start with having the same grade, but when the size of the class increases, the mid class tends to get decreasing grades, whilst the grades of the upperclass stays stable.