Indivisual Assignment 4

Data Visualization STAE04 (LU 22HT)

Yiying (Linda) Ren

2022-10-28

1 Task 1

In this assignment we will work with big and multivariate data.

The data set we will use for this assignment is the ´nlschools´ data set from the ´MASSLinks´ package. The data set features information on 11-year old (8th grade) pupils from 132 classes of 131 schools in the Netherlands and contains 2 287 observations and 6 variables. The variable of particular interest for us is *lang*, which represents test scores on a language test. Our goal is to see how this test score might be related to the other variables in the data set.

Source: Snijders, T. A. B. and Bosker, R. J. (1999) Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modelling. London: Sage.

1.1 Descriptive table of the variables

In Table 1, we outline the properties of the 6 variables in this data set, describing each in terms of its units of measurement and data type (continuous, integer, categorical). This data frame contains 2287 rows and the following columns.

Table 1: Description of the variables in the data set *nlschools* (eighth-Grade Pupils in the Netherlands from 1999).

Variable	Data Type	Class	Description
lang	discrete	integer	language test score
IQ	continuous	double	verbal IQ
class	discrete	factor	class ID.
GS	discrete	integer	class size
SES	discrete	integer	social-economic status of pupil's family
COMB	discrete	factor	Were the pupils taught in a multi-grade class (0/1)?

1.2 Data wrangling

Before we get started with visualizing the data set, we'll relabel the COMB variable such that the levels are more informative. Store the data set with a new name to

avoid over-writing the original data. We will also rename the variables to make them more informative.

First let's construct a naive visualization that examines the associations between class size, the type of class (single or multi-grade), and the score on the language test. Based on this simple first plots we will find better solutions to more clearly illustrate their relation with appropriate techniques.

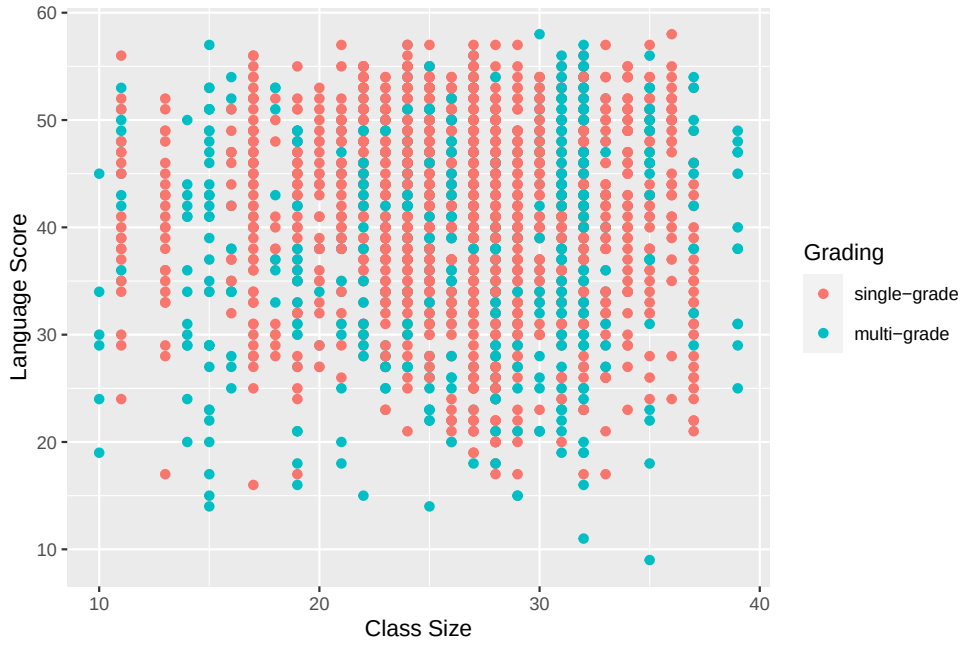

Figure 1: Naive Plot of language score vs size of the class grouped by the grading of class

As we can see in the naive plots of these three variables that it is a bit hard to see any pattern in this plot (although we could tell that majority of the students get graded by single-grade) hence for this naive plot to become a bit easier to read then we need to do further improvement.

We will use *smoothing* and *facet* to combat potential overlap with the aim of designing a visualization that clearly describes the relationship between these variables.

As we can see in the new scatter plot that using faceting method to individually illustrate the data with different grading systems helps to give us a less cluttered plot and with help with smoothing method, we can now easier tell the relationship between these variables.

First off, there are much more observations in single-grading and we can tell that majority of these 131 schools are very used to single-grading system; also schools who use single-grading seem to have higher language test score than schools who use multi-grading system in general; the language test score seem to increase as the class size increases but have a sudden decrease around class size of 24-27 and then starts to increase again as the class size increases for both grading systems (not sure why this happens); we can tell that the pattern between these variables doesn't seem to be linear and this might be due to that more variables from the data set should be

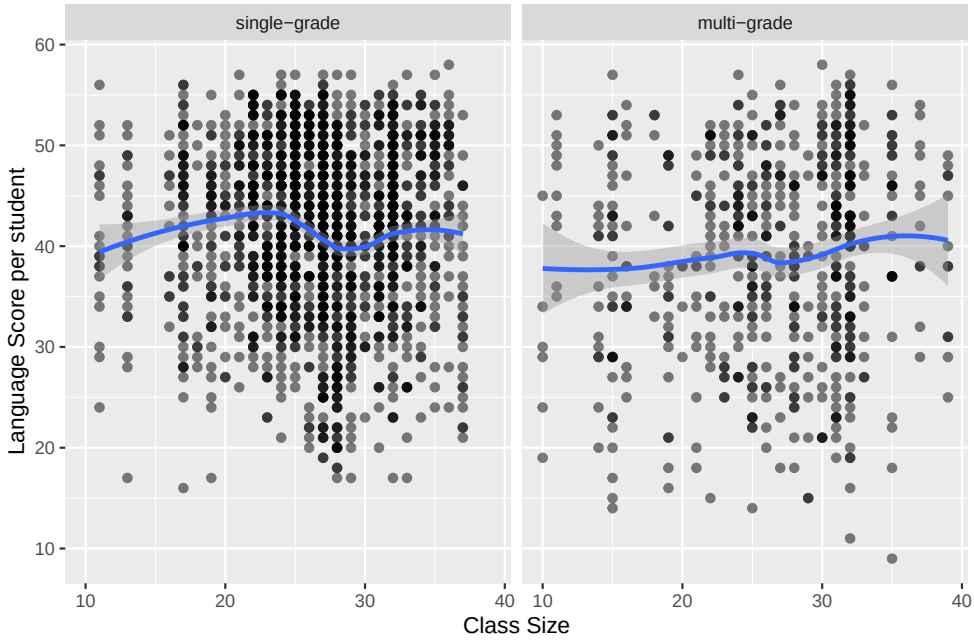

Figure 2: Scatter Plot using smoothing and facet method for language score vs size of the class, faceted by the grading of class

included to better explain the variable language test score. However as we can see there is an increasing trend in the pattern for both of the grading systems in spite of the clear sudden drop.

Ideally we'd probably like to see a straight horizontal line (based on average values of test score for different class sizes) for the test score across all class sizes since unbiased test scores in spite of class sizes is preferred. However the mean values of the class sizes for single-grading seem to be very fluctuated whereas it is much more steady for multi-grading, which is an indicator for its suitability for students in spite of the class size.

2 Task 2

One thing that the previous plot completely ignores is that the observations belong to different classes. This may be problematic because the scores among students in any given group reasonably are related because they are taught by the same teachers and so on. Hence we will use data wrangling skills to summarize the data set by computing the mean of language test scores inside each class. Then, we will produce a plot from the resulting data and compare the difference with the previous plot.

As we can see in the new cleaner plot that using the average score per class shrieked the data points to much fewer (132 classes in total) and three is much less overlapping and it becomes also easier to see the pattern in the plot. However the data still follows the same pattern as previous plot but just more enhanced in the curvatures due to different scales on the y-axis (ranging 20 - 50) since the previous density line was also based on the same value, e.g. the average language test score per class size.

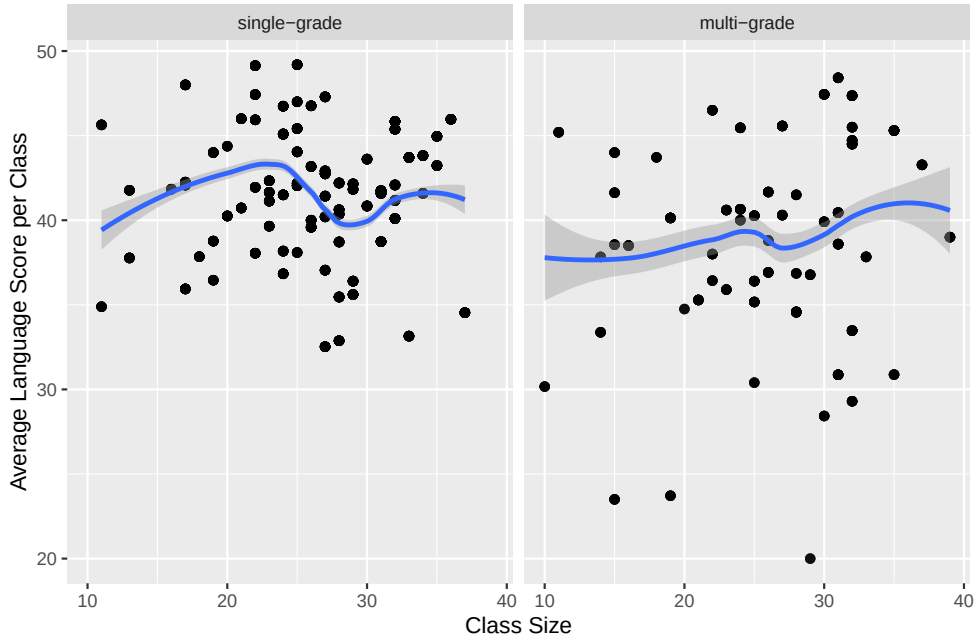

Figure 3: Plots for the Average Language Score vs Size of the Class grouped by the Grading of class

3 Task 3

For the final task, we will create one more plot including another variable *social economical status* (SES) as well. We will be working on the original data set with the individual test score per student and in order to make the illustration work well, we will convert variable *class size* (size) from a numerical variable into a factor variable with several levels. To summarize, we will create a plot that includes all of the 4 variables: *language test score* (lang), *class size* (GS), *social economical status* (SES), and *grading type* (COMB).

In order to make a nice plot with 4 variables, we could convert *size* from a numerical variable into a factor variable with several levels, which we can then use to group the data with. It seems just reasonable to cut the levels by 5 (with consideration to its range from 10 to 39) between class sizes (of course we could also do it by 3-4 to get even more levels in attempt for revealing more details in the plot). Now with this additional information on social economical status included in the plot, we could see that in spite of class sizes, it seems that students with higher social economic status background tend to have higher test scores for both grading systems. Now with taking social economical status into account, now we can see that we could no longer say that the average language test score tends to fluctuate for single grading system much more than for the multi-grading system (grading bias) alongside the increase of class size. There might be more factors (variable such as verbal IQ) in the data set that need to be included to reveal more co-relations to the language test scores, which is out of the scope of this assignment. But if we are to investigate further with verbal IQ included, we could then first convert social economical status into a factor with several levels (preferably 3 levels: low, medium and high) and plot

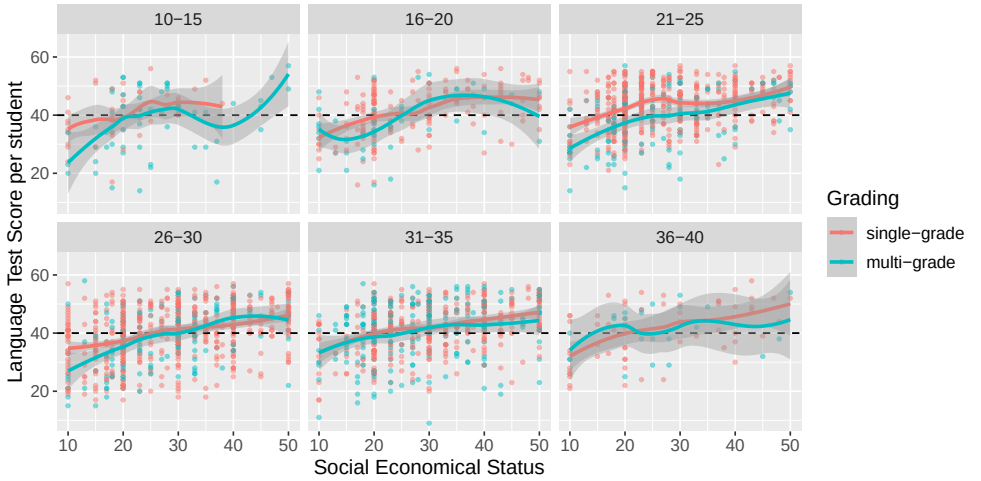

Figure 4: Scatter Plot of Individual Language Score vs the Social Economical Status, grouped by Size of the Class and the Grading of Class, with converting class size into factor with 6 levels

verbal IQ vs language test score with faceting both class size and social economical status and color the observation by grading.

As a short summary of this assignment, we could say that our findings are students who have higher social economical status tend to have higher test scores in spite of the class size and grading system.