Assignment 4

STAE04 - Statistics: Data Visualisation

Thi Kim Hong Nguyen

2024-02-28

In this assignment, we'll work with big and multivariate data. We will use the nlschools data set from MASS package, was used in 1999 by Snijders and Bosker as an example in their Multilevel Analysis work.

1 Task 1: Getting the Data

The data set comprises 2287 observations and encompasses six variables. These variables were collected as part of a study involving eighth-grade pupils, approximately 11 years old, across 131 schools in the Netherlands. The variables can be described as follows:

Table 1: The variables of the data set

Attribute	Description
lang	language test score
IQ	verbal IQ
class	class ID
GS	class size: number of eighth-grade pupils recorded in the class
SES	social-economic status of pupil's family
COMB	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

Prior to conducting any additional analysis, some of the variables were renamed to make the interpretation process easier. 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. COMB has been renamed to Class.type, GS to Class.size, lang to Test.score, and class to Class.Id. We are particularly interested in the variable "lang," which signifies test scores on a language test. Our objective is to explore the potential associations between this test score and the other variables within the data set. Let's take a look at the data set after renaming

```
## Rows: 2,287
## Columns: 6
## $ Test.score <int> 46, 45, 33, 46, 20, 30, 30, 57, 36, 36, 29~
## $ IQ         <dbl> 15.0, 14.5, 9.5, 11.0, 8.0, 9.5, 9.5, 13.0~
```

```
## $ Class.Id   <fct> 180, 180, 180, 180, 180, 180, 180, 180, 18~
## $ Class.size <int> 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29~
## $ SES        <int> 23, 10, 15, 23, 10, 10, 23, 10, 13, 15, 10~
## $ Class.type <fct> single-grade, single-grade, single-grade, ~
```

```
# Create a scatter plot
ggplot(nlschools_renamed, aes(Class.size, Test.score, color = as.factor(Class.type))) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "loess") +
  facet_wrap(~ Class.type) +
  labs(x = "Class Size",
       y = "Language Test Score",
       color = "Class Type")
```

Figure 1: Associations between Class Size, Class Type, and Language Test Score

This visualization indicates that multi-grade classes generally accommodate a larger number of students, and their average grades are only slightly lower than those in the other group. One possible inference from this observation is that smaller classes might contribute to better academic performance, as teachers can provide more focused attention to each student. Notably, single-grade classes demonstrate improved performance with an optimal class size of around 25, while multi-grade classes show optimal performance with sizes slightly above 30. To some extent, one could suggest a modest correlation between mid-sized single-grade classes and higher academic achievement.

2 Task 2: Exploring the Depths of the Data

The previous plot overlooks an essential aspect: the observations belong to distinct classes. This oversight may pose challenges as the scores among students within the

same group are likely correlated, given that they share common teachers and other factors. I will summarize the data set by computing the mean or median (your choice) language test scores inside each class.

```
# Summarize the data by computing the mean language test score for each class
summary_by_class <- nlschools_renamed %>%
  group_by(Class.Id, Class.size, Class.type) %>%
  summarize(Mean_Test_Score = mean(Test.score, na.rm = TRUE))

# View the summary data
print(summary_by_class)
```

```
## # A tibble: 133 x 4
## # Groups:   Class.Id, Class.size [133]
##    Class.Id Class.size Class.type   Mean_Test_Score
##    <fct>         <int> <fct>                  <dbl>
##  1 180              29 single-grade            36.4
##  2 280              19 multi-grade             23.7
##  3 1082             25 multi-grade             30.4
##  4 1280             31 multi-grade             30.9
##  5 1580             35 multi-grade             30.9
##  6 1680             28 multi-grade             41.5
##  7 1880             28 single-grade            32.9
##  8 2180             19 single-grade            38.8
##  9 2480             35 single-grade            45.0
## 10 2680             21 single-grade            40.7
## # i 123 more rows
```

```
# Plot the relationship between class size, class type, and mean language test score
ggplot(summary_by_class, aes(Class.size, y = Mean_Test_Score, color = as.factor(Class.type))) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(x = "Class Size",
       y = "Mean Language Test Score",
       color = "Class Type")
```

This plot provides a more generalized overview, suitable for identifying overall trends and comparing average performance across different class types. However, the conclusion is still the same with the previous plot. However, summarize plot may not capture the full distribution of scores within each class. Outliers or variations may not be fully represented, it may also oversimplify the diversity that can exist within class types. To mitigate these disadvantages, it's crucial to carefully choose the summarization approach based on the research question and data set characteristics. It's recommended to thoroughly compare the insights gained from both the original and summarized data sets, keeping in mind the trade-offs and limitations associated with data sumarization.

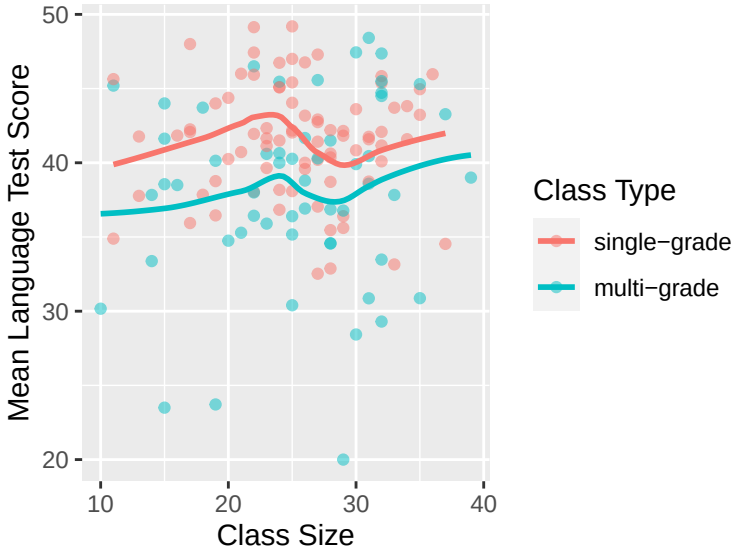

Figure 2: Relationship between Class Size, Class Type, and Mean Language Test Score

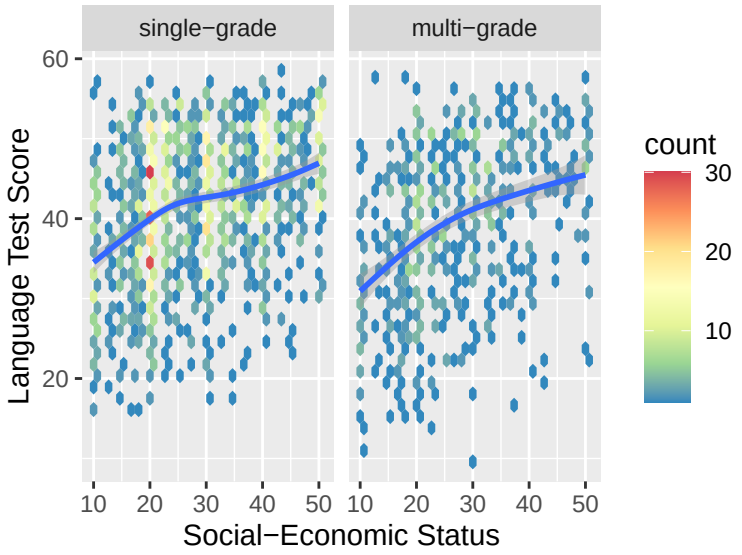

Figure 3: Effecy of socio-economic status, Class size, Class type on Language Test Score

3 Task 3: Exploring the Depths of the Data

As anticipated, a distinct correlation emerges among these variables. A positive association is evident between socio-economic status and language test scores, which is quite reasonable given that a higher socio-economic status often implies better educational opportunities, increased access to resources like food and leisure, and other factors that can influence test performance. Additionally, this plot underscores the presence of inequality. The majority of children fall within the socio-economic status range of 10-25. The positively skewed distribution may suggest a prevalence of individuals in lower socio-economic levels, or it could reflect the possibility that the studied schools predominantly represented these socio-economic statuses.

In summary, the significant advantage of this particular plot, as compared to others, lies in its ability to carve out specific segments in the data for independent comparisons. While it suggests that socio-economic status tends to elevate the overall score, the process of testing various methods and selecting specific elements to enhance clarity in the plot can be quite cumbersome. In retrospect, it might have been more straightforward to present several visualizations highlighting the most critical variables, as class size, for instance, may not appear as crucial as initially thought.