ANLYTC3 Mini-Project

### Instructions

The data set nlschools (in the MASS library, use data(nlschools, package = "MASS") to load in the data) contains the following records for 2287 students in the Netherlands (see ?nlschools for more information)

* lang: their test score on a language exam
* IQ: their verbal IQ
* class: the ID number for their classroom
* GS: the number of students in each class
* SES: the socio-economic status of their family
* COMB: was the student in a multi-grade class? (0=no, 1=yes)

Your task is to explore and describe this data set. You should use both models and visualization. You should use your judgement as to what are the most important aspects of the data to highlight; however, the following questions are of particular interest:

1. Are there discrepancies in IQ or SES in the different classes, or when grouping by multi-grade vs non-multi-grade classes?
2. When did students perform better or worse on the language exam? Describe which variables had the most important effects.
3. Do you think there are interactions in the effects of the variables on the language exam score? Speculate as to the cause of any such effects that you think should be included.

Always think about what you are doing – blindly replicating the lectures may not result in the best analysis! Every data set is different, so you may need to use your judgement on the right thing to do. Your goal is to understand the data as best as your can, particularly in regards to questions 1-3 above.

NOTE: Since this is an observational study, we’ll understand that any claims about causality should be taken mainly as speculation or hypothesis. Usage of language regarding the “effects” of a variable on the outcome of interest will be understood to be informal, so we won’t penalize you. (This will probably make the report much easier to write.)

### Presentation Requirements

Your output should be this rmarkdown file, which should compile into a report that gives your analysis.

The report should be understandable. One possible outline is:

1. Intro: what are your main steps, and what are your findings?
2. Analysis: Describe each step and why you did it, show the R code and results, then comment on the results.
3. Conclusions: recap and summarize

Typically, it’s best to do step 2 first, and then write step 1 and 3 after you are finished. Bulleted lists may be especially helpful for these parts.

Code should be clearly commented. All plots should be presentable and properly labeled. Mitigate overplotting whenever possible. When building a model, explain each step so that we know why you are doing it.

The due date is 11:59pm EST June 28, 2019 (Friday). This is a hard deadline – no late days may be used. Submit this Rmd file on Canvas Don’t submit additional files.

### Report:

#### Introduction

The dataset that the Analytics Practitioner will be working on is the nlschools dataset which is a study of 2,287 eighth-grade pupils (aged about 11) in 132 classes in 131 schools in the Netherlands conducted by Snijders and Bosker (1999).

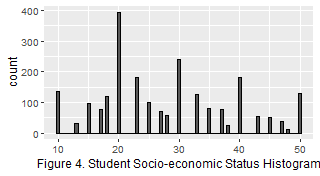
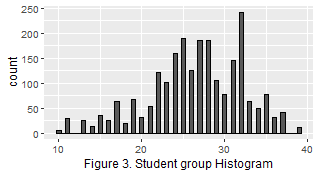
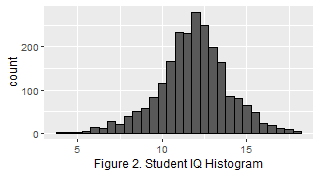
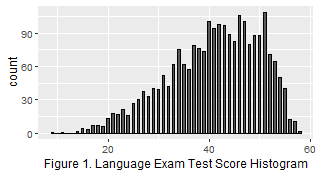
To get a glimpse of what the dataset is about, the Analytics Practitioner determined the average values of some attributes as detailed below:

1. The dataset has 2287 rows.
2. The average language exam test score of all the students is 40.9348491.
3. The standard deviation of the students’ language exam is 9.003676.

The below tables show the attribute summary of the nlschools dataset.

Table 1: Summary of the Dataset

|  | lang | IQ | class | GS | SES | COMB |
| --- | --- | --- | --- | --- | --- | --- |
|  | Min. : 9.00 | Min. : 4.00 | 15580 : 33 | Min. :10.00 | Min. :10.00 | 0:1658 |
|  | 1st Qu.:35.00 | 1st Qu.:10.50 | 5480 : 31 | 1st Qu.:23.00 | 1st Qu.:20.00 | 1: 629 |
|  | Median :42.00 | Median :12.00 | 15980 : 31 | Median :27.00 | Median :27.00 | NA |
|  | Mean :40.93 | Mean :11.83 | 16180 : 31 | Mean :26.51 | Mean :27.81 | NA |
|  | 3rd Qu.:48.00 | 3rd Qu.:13.00 | 18380 : 31 | 3rd Qu.:31.00 | 3rd Qu.:35.00 | NA |
|  | Max. :58.00 | Max. :18.00 | 5580 : 30 | Max. :39.00 | Max. :50.00 | NA |
|  | NA | NA | (Other):2100 | NA | NA | NA |



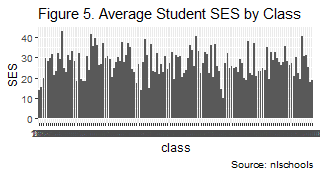


Table 1 and Figures 1-5 show the summary of the dataset which details the following:

1. The highest language exam score of a student (lang column) is 58, while its average is 40.93, and the lowest is 9.
2. The highest student IQ (IQ column) is 18, while its average is 11.83, and the lowest is 4.
3. The highest student count (GS column) in a class is 39, while its average is 26.51, and the lowest is 10.
4. The highest socio-economic status (SES column) of a student is 50, while the average is 27.81, and the lowest is 10.
5. There are a total of 629 students who are under a multi-grade class (COMB column), while there are 1,658 students who are not.

#### Body

To expound on the summary of the dataset, the Analytics Practitioner will answer the questions posed using tables and graphs.

**1. Are there discrepancies in IQ or SES in the different classes, or when grouping by multi-grade vs non-multi-grade classes?**

Table 2: The classes with the highest standard deviation in IQ

| sd.IQ | avg.IQ | class | COMB | GS | avg.lang | sd.lang | avg.SES | sd.SES |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 3.52 | 9.75 | 12380 | 1 | 20 | 34.75 | 4.27 | 24.00 | 11.46 |
| 3.42 | 13.40 | 2980 | 1 | 22 | 46.50 | 6.22 | 42.70 | 6.02 |
| 2.97 | 10.50 | 21480 | 0 | 27 | 32.52 | 8.58 | 23.05 | 10.15 |
| 2.89 | 12.77 | 21880 | 0 | 27 | 47.29 | 5.50 | 35.04 | 9.00 |
| 2.81 | 12.12 | 13080 | 1 | 24 | 45.46 | 9.47 | 19.15 | 7.17 |
| 2.69 | 11.76 | 16080 | 0 | 26 | 43.17 | 9.52 | 32.26 | 15.53 |
| 2.67 | 10.82 | 10980 | 0 | 33 | 33.15 | 6.98 | 14.75 | 4.83 |
| 2.56 | 10.80 | 10180 | 0 | 25 | 42.17 | 8.33 | 26.74 | 6.94 |
| 2.56 | 10.32 | 180 | 0 | 29 | 36.40 | 8.76 | 13.84 | 4.70 |
| 2.50 | 6.75 | 10380 | 1 | 29 | 20.00 | 6.63 | 13.75 | 4.79 |

As seen in Table 2, the highest standard deviation in the students’ IQ is class 12380 with 3.52 standard deviation and an average language score of 34.75. The second class with the highest standard deviation is class 2980 with 3.42 standard deviation and 46.50 average language score. As seen in Table 2, the top 2 classes with the highest deviation in IQ are those in multi-grade classes.

Table 3: The classes with the lowest standard deviation in IQ

| sd.IQ | avg.IQ | class | COMB | GS | avg.lang | sd.lang | avg.SES | sd.SES |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.84 | 11.69 | 1680 | 1 | 28 | 41.50 | 6.30 | 29.38 | 5.29 |
| 0.88 | 11.97 | 9080 | 0 | 22 | 45.93 | 5.20 | 36.67 | 9.76 |
| 0.90 | 12.56 | 21780 | 0 | 20 | 44.38 | 6.16 | 23.88 | 7.43 |
| 0.90 | 12.77 | 20680 | 0 | 11 | 45.64 | 6.04 | 22.27 | 5.27 |
| 0.92 | 11.20 | 4480 | 1 | 25 | 36.40 | 8.87 | 19.27 | 4.27 |
| 1.07 | 12.88 | 13680 | 1 | 30 | 47.44 | 5.82 | 30.06 | 9.31 |
| 1.16 | 10.75 | 19280 | 1 | 14 | 33.38 | 6.72 | 24.88 | 9.16 |
| 1.19 | 11.00 | 18880 | 1 | 26 | 36.92 | 9.67 | 24.83 | 8.38 |
| 1.19 | 12.11 | 11880 | 1 | 29 | 36.78 | 8.29 | 26.78 | 9.74 |
| 1.21 | 11.33 | 3380 | 1 | 14 | 37.83 | 10.65 | 24.50 | 7.42 |

As seen in Table 3, the class with the lowest standard deviation in IQ is class 1680 which is a multi-grade class and has an average IQ of 11.69. It also has an average language score of 41.50.

It can be inferred in Tables 2 and 3 that the lower the standard deviation in IQ is, the higher the language test score of the students are. This can be seen in the Table 3 which shows the classes having the lowest standard deviation and the said table shows high test scores.

Table 4: The classes with the highest standard deviation in SES

| sd.SES | avg.SES | class | COMB | GS | avg.lang | sd.lang | avg.IQ | sd.IQ |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 15.53 | 32.26 | 16080 | 0 | 26 | 43.17 | 9.52 | 11.76 | 2.69 |
| 15.17 | 30.33 | 4980 | 1 | 28 | 34.56 | 11.95 | 10.94 | 2.13 |
| 15.01 | 28.00 | 23580 | 1 | 15 | 44.00 | 9.19 | 12.05 | 1.38 |
| 14.64 | 32.50 | 22680 | 1 | 24 | 40.00 | 7.69 | 12.06 | 1.84 |
| 14.07 | 26.33 | 23780 | 1 | 19 | 40.13 | 7.01 | 11.17 | 2.16 |
| 13.88 | 27.86 | 8080 | 0 | 28 | 40.62 | 9.97 | 10.83 | 2.06 |
| 13.14 | 32.11 | 18480 | 0 | 29 | 42.14 | 7.44 | 12.48 | 1.96 |
| 12.44 | 36.59 | 11080 | 0 | 19 | 44.00 | 5.00 | 12.24 | 1.30 |
| 12.21 | 21.13 | 20980 | 0 | 27 | 42.91 | 7.06 | 12.07 | 1.32 |
| 12.15 | 33.20 | 15080 | 0 | 30 | 40.84 | 7.46 | 11.56 | 1.23 |

Table 4 shows the classes with the highest standard deviation in socio-economic status (SES) of the students. The highest standard deviation in SES is class 16080 which has an average SES of 32.26 and is not a multi-grade class.

Table 5: The classes with the lowest standard deviation in SES

| sd.SES | avg.SES | class | COMB | GS | avg.lang | sd.lang | avg.IQ | sd.IQ |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.00 | 10.00 | 18280 | 0 | 11 | 34.89 | 6.64 | 9.22 | 2.09 |
| 2.39 | 22.38 | 24380 | 1 | 15 | 41.62 | 7.17 | 10.38 | 1.33 |
| 3.42 | 21.24 | 2180 | 0 | 19 | 38.76 | 9.73 | 11.76 | 2.09 |
| 3.90 | 20.15 | 6880 | 0 | 23 | 41.65 | 8.00 | 12.85 | 2.10 |
| 3.95 | 19.15 | 24480 | 0 | 13 | 37.77 | 9.98 | 11.12 | 2.19 |
| 4.27 | 19.27 | 4480 | 1 | 25 | 36.40 | 8.87 | 11.20 | 0.92 |
| 4.32 | 21.96 | 16480 | 0 | 25 | 44.04 | 6.62 | 11.73 | 1.59 |
| 4.51 | 20.56 | 21280 | 0 | 25 | 42.04 | 7.36 | 11.94 | 2.34 |
| 4.70 | 13.84 | 180 | 0 | 29 | 36.40 | 8.76 | 10.32 | 2.56 |
| 4.79 | 13.75 | 10380 | 1 | 29 | 20.00 | 6.63 | 6.75 | 2.50 |

Table 5 shows the classes with the lowest standard deviation in SES. Class 18280 has 0 SES, which is the lowest one. The said class has an average SES of 10 and has 11 students under the class.

It can be inferred in Tables 4 and 5 that the higher the standard deviation in SES is, the higher the average language test score of the class is.

Table 6: The classes with the highest language test scores

| avg.lang | sd.lang | class | COMB | GS | avg.IQ | sd.IQ | avg.SES | sd.SES |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 49.19 | 4.93 | 22880 | 0 | 25 | 12.36 | 2.09 | 27.33 | 6.77 |
| 49.14 | 5.25 | 14780 | 0 | 22 | 12.66 | 1.24 | 23.50 | 4.82 |
| 48.42 | 6.19 | 19380 | 1 | 31 | 12.67 | 1.54 | 25.00 | 8.45 |
| 48.00 | 5.65 | 15280 | 0 | 17 | 12.38 | 1.24 | 40.24 | 6.84 |
| 47.44 | 5.82 | 13680 | 1 | 30 | 12.88 | 1.07 | 30.06 | 9.31 |
| 47.43 | 6.40 | 13281 | 0 | 22 | 12.74 | 1.50 | 30.86 | 10.71 |
| 47.36 | 6.58 | 6580 | 1 | 32 | 12.57 | 1.64 | 29.77 | 11.77 |
| 47.29 | 5.50 | 21880 | 0 | 27 | 12.77 | 2.89 | 35.04 | 9.00 |
| 47.00 | 6.13 | 9480 | 0 | 25 | 13.73 | 2.40 | 34.73 | 9.88 |
| 46.76 | 5.22 | 6280 | 0 | 26 | 13.64 | 2.15 | 36.67 | 10.53 |

Table 6 shows the classes with the highest language scores. It can be seen that class 22880 has the highest average language test score and has 4.93 standard deviation in the language test scores. The table also shows that the the class is not a multi-grade class.

It can be inferred in Table 6 that the classes with the highest test scores have student counts not lower than 15.

Table 7: The classes with the lowest language test scores

| avg.lang | sd.lang | class | COMB | GS | avg.IQ | sd.IQ | avg.SES | sd.SES |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 20.00 | 6.63 | 10380 | 1 | 29 | 6.75 | 2.50 | 13.75 | 4.79 |
| 23.50 | 4.78 | 4780 | 1 | 15 | 8.31 | 2.39 | 18.12 | 5.79 |
| 23.71 | 6.32 | 280 | 1 | 19 | 9.00 | 2.10 | 15.43 | 4.79 |
| 28.43 | 6.97 | 25880 | 1 | 30 | 10.64 | 1.28 | 19.00 | 6.19 |
| 29.30 | 13.28 | 25680 | 1 | 32 | 10.55 | 2.33 | 17.80 | 8.35 |
| 30.17 | 8.93 | 17980 | 1 | 10 | 11.25 | 1.33 | 14.50 | 5.05 |
| 30.40 | 10.16 | 1082 | 1 | 25 | 10.50 | 1.62 | 19.60 | 8.56 |
| 30.87 | 6.97 | 1280 | 1 | 31 | 9.43 | 1.25 | 29.47 | 9.46 |
| 30.88 | 13.30 | 1580 | 1 | 35 | 11.50 | 2.38 | 28.12 | 8.43 |
| 32.52 | 8.58 | 21480 | 0 | 27 | 10.50 | 2.97 | 23.05 | 10.15 |

Table 7 shows the classes with the lowest language scores. It can be seen that class 10380 has the lowest language test score average and the said class is multi-graded. The average IQ of the students under the class is 6.75 which is lower than 11.83–the overall average of the dataset.

It can be inferred from the table that the students enrolled in multi-grade class has lower language test scores.

**2. When did students perform better or worse on the language exam? Describe which variables had the most important effects.**

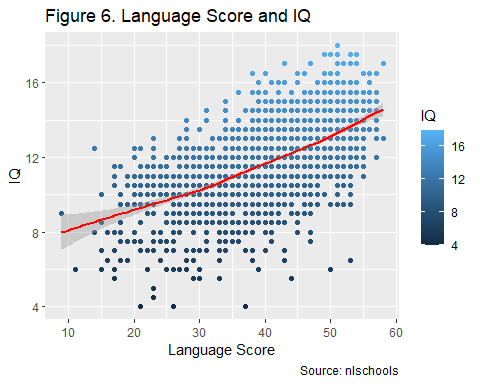


Figure 6 shows that the higher the IQ of the student is, the higher the student’s language test score is. Meaning, the student IQ is directly proportional with their language test score.

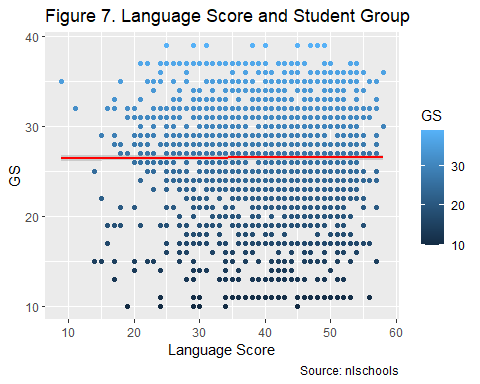


Figure 7 shows that there is not much change in the language test score if the group of students is increased in a class.

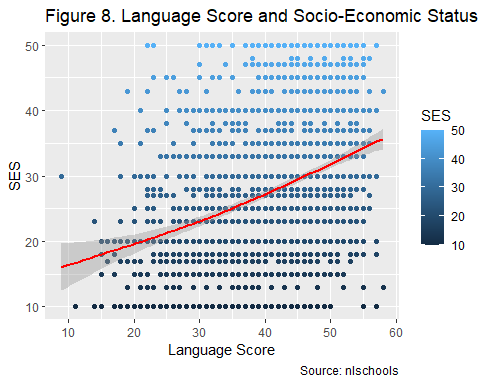


Figure 8 shows that the higher the socio-economic status of the student is, the higher the language test score is. Hence, the socio-economic status (SES) is directly proportial with the language test score.

Table 8: The average language score by COMB

| COMB | avg.lang | sd.lang | avg.IQ | sd.IQ | avg.SES | sd.ses | count.lang |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 41.60 | 8.51 | 11.90 | 2.05 | 28.13 | 11.01 | 1658 |
| 1 | 39.18 | 10.00 | 11.66 | 2.11 | 26.98 | 10.60 | 629 |

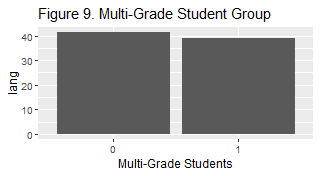
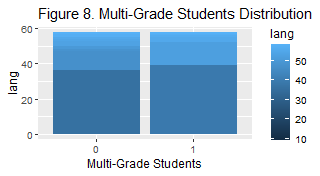


Table 8, and Figures 8 and 9 show that the students under the non-multigrade class have higher average language test scores and lower standard deviation in their language test scores. It can also be seen that the non-multigrade class has lower count of students with low test scores compared with those in multi-grade class as shown in the shade of blue in Figure 8.

**3. Do you think there are interactions in the effects of the variables on the language exam score? Speculate as to the cause of any such effects that you think should be included.**

Based on what the graphs infer, yes, there are interactions between the variables. Some variables that have directly proportional impact on the language exam test score are the following:

1. Student IQ - As seen on Figures 6, the language exam test score is directly proportional to the IQ of the student. This only means that IQ is the highest indicator of the language exam test score in the dataset. This is expected since a student’s IQ is a big indicator of a student’s test score in most subjects.
2. Socio-economic status of the student - As seen on Figure 8, the socio-economic status of the student has direct correlation to the student’s language exam test score. The Analytics Practitioner speculates that the richer the family of the student is, the more guidance and supplemental materials can be purchased by the family to aid the student’s studies, particularly in language-related subjects
3. COMB (whether the student is in multi-grade or not) - As seen in Tables 8 and Figures 8 and 9, students enrolled in a multi-grade classes have lower language exam test scores compared to those who are not enrolled in a multi-grade class. It can be inferred from the graphs that the students are having difficulty maximizing their potential in the language subject if the teacher is handling multi-grade classes, which impacts the student’s language test scores.

#### Conclusion

As a conclusion, the following can be inferred from the dataset:

1. The IQ of the students enrolled in a class are directly proportional with their language test scores.
2. Those with high socio-economic status have higher language test scores.

To maximize the language test scores of the students, the school can take the following action plans:

1. Give scholarships to students with High IQ but low to average socio-economic status.
2. Minimize or remove multi-grade classes from the curriculum so that the teacher can focus on a specific student group’s development levels and needs. This is because the data shows that the students find it difficult to maximize their language test scores if they are enrolled in a multi-grade classes.