What distinguishes high-performing students?

Will Landau

April 25, 2015

Introduction

The Data

In 2012, the Organization for Economic Co-operation and Development (OECD) Programme for International Student Assessment (PISA) surveyed roughly five hundred thousand, fifteen-year-old students from sixty-five economies across the globe ("Organization for Economic Co-operation and Development" 2015). Questions measured students' reading, math, and science skills in ways that, according to the OECD website, "are not directly linked to the school curriculum. The tests are designed to assess to what extent students at the end of compulsory education, can apply their knowledge to real-life situations and be equipped for full participation in society" ("Organization for Economic Co-operation and Development" 2015). Students also answered extensive background questionnaires about their study habits, attitudes towards school, circumstances at home, etc. Extensive data were recorded about the schools and parents of those students as well.

PISA math and reading scores naturally divide USA subjects into high-performing students (top 25%) and low-performing students (bottom 25%). I select a subset of variables in the student-specific PISA data and attempt to predict student performance (high or low) in the United States.

Motivating questions

- In general, which issues are most strongly associated with student success in the USA?
- What kinds of students are there? If they partition into groups, what do these groups mean?
- How strongly to the PISA metrics predict student success?

Exploratory analysis, variable selection, and preprocessing

The need for subsetting students and censoring responses

Table 1 demonstrates the need to subset the USA data and censor the reading and math scores. All the scores are tightly correlated, so predicting them as continuous responses may not be productive, especially since most of the student data is categorical. However, there is a way to make the responses more manageable. First, I remove the middle 50% of students (students for whom the sum of all the reading and math scores is between the 25th and 75th percentile). Second, I censor the data: students scoring above the 75th percentile overall are designated high-performing, and those scoring below the 25th percentile are designated low-performing.

Table 1: correlation matrix of the raw reading and math scores of USA students. All the scores are tightly correlated, so predicting them as continuous responses may not be productive, especially since most of the student data is categorical. However, there is a way to make the responses more manageable. First, I remove the middle 50% of students (students for whom the sum of all the reading and math scores is between the 25th and 75th percentile). Second, I censor the data: students scoring above the 75th percentile overall are designated high-performing, and those scoring below the 25th percentile are designated low-performing.

	mth1	rd1	mth2	rd2	mth3	rd3	mth4	rd4	mth5	rd5
PV1MATH	1.00	0.87	0.93	0.83	0.93	0.83	0.93	0.82	0.93	0.83
PV1READ	0.87	1.00	0.83	0.90	0.83	0.90	0.83	0.90	0.83	0.91
PV2MATH	0.93	0.83	1.00	0.87	0.93	0.83	0.93	0.83	0.93	0.83
PV2READ	0.83	0.90	0.87	1.00	0.83	0.90	0.83	0.90	0.83	0.90
PV3MATH	0.93	0.83	0.93	0.83	1.00	0.87	0.93	0.83	0.93	0.83
PV3READ	0.83	0.90	0.83	0.90	0.87	1.00	0.83	0.90	0.83	0.90
PV4MATH	0.93	0.83	0.93	0.83	0.93	0.83	1.00	0.87	0.93	0.83
PV4READ	0.82	0.90	0.83	0.90	0.83	0.90	0.87	1.00	0.82	0.90
PV5MATH	0.93	0.83	0.93	0.83	0.93	0.83	0.93	0.82	1.00	0.87
PV5READ	0.83	0.91	0.83	0.90	0.83	0.90	0.83	0.90	0.87	1.00

Finding useful predictor variables using a matching heuristic

The student-specific USA data has roughly 500 variables for predicting reading and math scores. I remove the few continuous variables and the variables with all missing values. From the 256 remaining categorical variables, I remove the ones related to self-efficacy, familiarity, and experience with math and reading concepts. (Predicting on these would be logically circular, nearly amounting to cheating.) In addition, I remove a few nondescript "ISCED" variables with poor documentation. I am left with 210 factor variables for prediction.

To comb through the remaining 210 factor variables efficiently, I make use of a simple matching heuristic. The heuristic is similar to the simple matching coefficient for comparing two categorical variables, and it loosely measures how well a factor accurately splits students according to high or low performance. For each factor x, I calculate this version of the matching coefficient as follows.

1. Remove the missing values from x, along with the corresponding values from the binary vector y of student performances (high and low coded as 1 and 0, respectively).
2. For every subset s of the levels of $x = (x_1, \ldots, x_n)$:
   a. Create the binary vector $z = (z_1, \ldots, z_n)$, where $z_i = I(x_i \in s)$.
   b. Let the matching score of s be

$$\frac{1}{n} \max \left\{ \sum_{i=1}^{n} I(y_i = z_i),\ \sum_{i=1}^{n} I(y_i \neq z_i) \right\}$$

3. Take the matching score of x to be the maximum of all the matching scores calculated in step 2.

This matching score is the most optimistic rate of correct classification for a prediction on a single variable. A matching of 1 means that x can predict y perfectly, and a rate of 0.5 means that x is no better than chance.

Figure 1 shows the matching scores of the 210 candidate factors. Most variables are better than chance. The two variables with matchings better than 0.7 are "How many books at home" (censored) and "Vignette Classroom Management - Students Frequently Interrupt/Teacher Arrives Late".

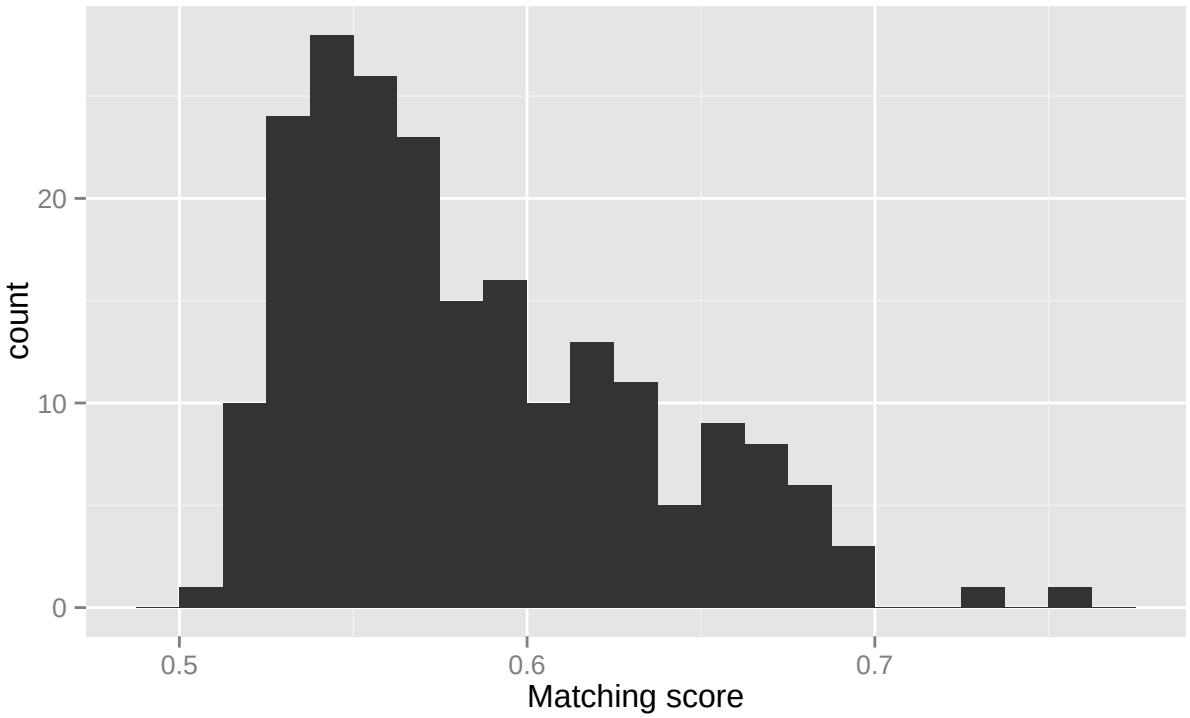

Figure 1: matching scores of all 210 candidate factors. Most variables are better than chance. The two variables with matchings better than 0.7 are "How many books at home" (censored) and "Vignette Classroom Management - Students Frequently Interrupt/Teacher Arrives Late".

Figure 2 shows the matching scores of the 210 variables, but this time, the variables are grouped by the general issues they cover, such as possessions, attitude, and parental backgrounds. Here, we can get a general picture of which issues matter in the sample of USA students. Academic habits outside school, school-related possessions, and attendance/truancy seem to be most related to performance, although there are few variables on these topics. There is a lot of variety among the attitude and teaching variables, and most of the survey questions seem to cover these issues. Many attitude and teaching questions are important, and will be used in the classifiers developed later. Interestingly, gender, course-content, and sociality (variables that attempt to measure the quality of students' social lives and social norms about school) were not very important. The parent-related variables seem largely split two ways, but this split does not separate mother-related variables from father-related ones.

Variable selection and imputation

Figure 3 shows the variables with the top 20 matching scores calculated in the previous section. I will use these to build a classifier.

Unfortunately, even with a reduction in the number of variables, there are still a lot of missing values. Figure 4 show the number of missing values for each student, and Figure 5 shows the percentage of missing values within each variable. The figures lead to the following imputation strategy:

1. Remove the students who missed more than 13 questions (only 3.4% of USA students). The resulting dataset 5052 students and no missing values. In Figure 6, I determine that no more students need to be removed for the imputation to be reasonable.
2. Put all factors on a numeric scale such that the natural ordering of factor levels is preserved. (All factors are ordinal.) Center and scale the predictor variables, and denote student performance by 1 and -1 for high-performing and low-performing students, respectively.

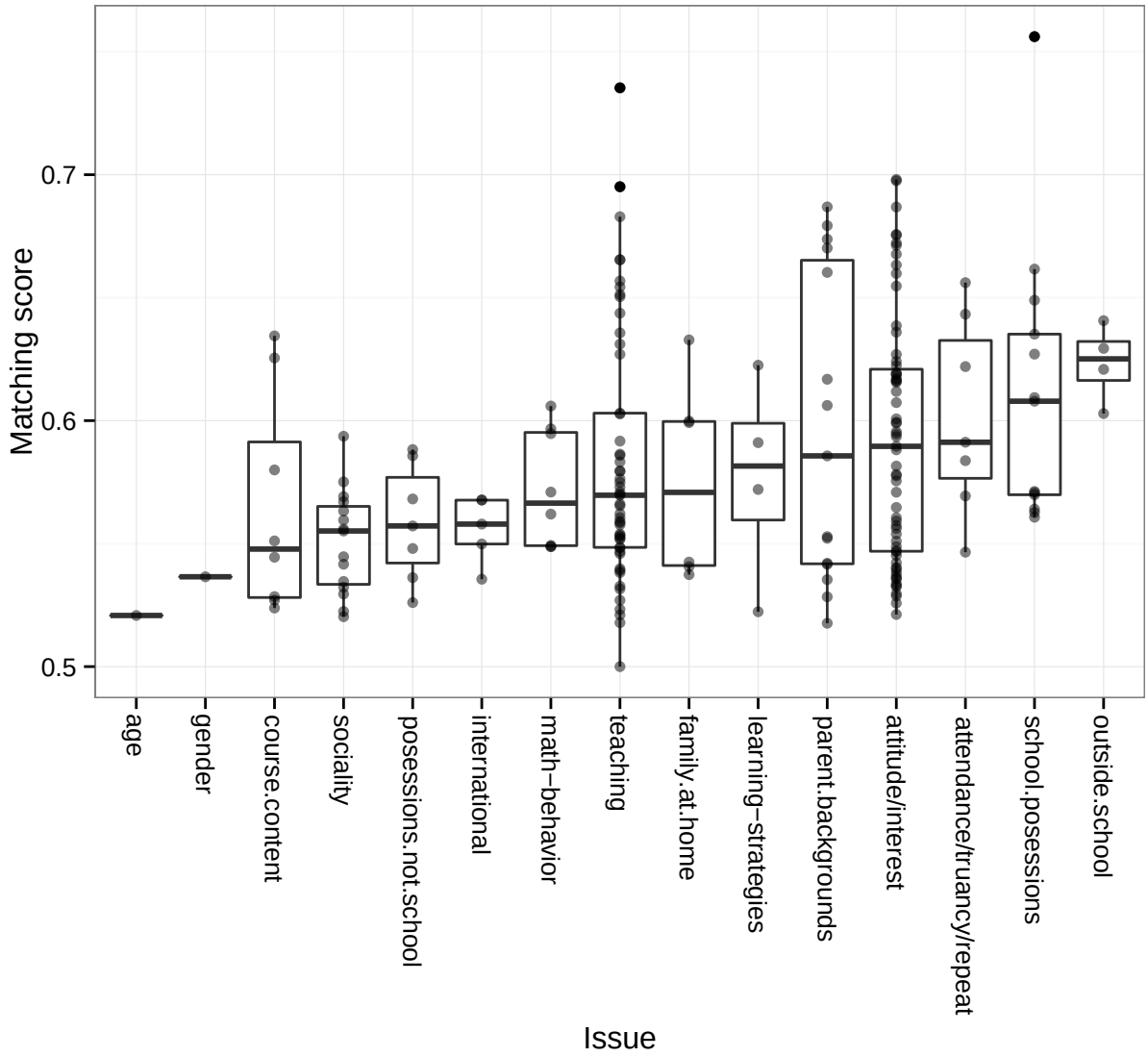

Figure 2: matching scores of the 210 variables, but this time, with the variables grouped by the general issues they cover, such as possessions, attitude, and parental backgrounds. Here, we can get a general picture of which issues matter in the sample of USA students. Academic habits outside school, school-related possessions, and attendance/truancy seem to be most related to performance, although there are few variables on these topics. There is a lot of variety among the attitude and teaching variables, and most of the survey questions seem to cover these issues. Many attitude and teaching questions are important, and will be used in the classifiers developed later. Interestingly, gender, course-content, and sociality (variables that attempt to measure the quality of students' social lives and social norms about school) were not very important. The parent-related variables seem largely split two ways, but this split does not separate mother-related variables from father-related ones.

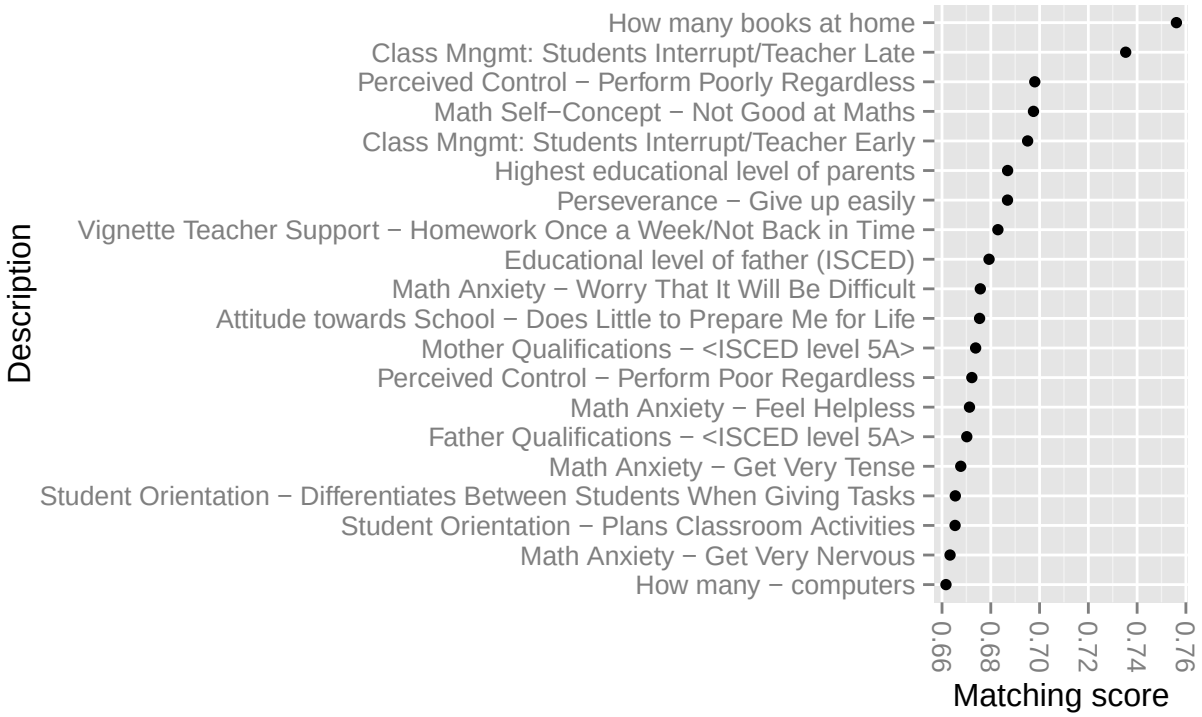

Figure 3: variables with the top 20 matching scores. I will use these to build classifiers.

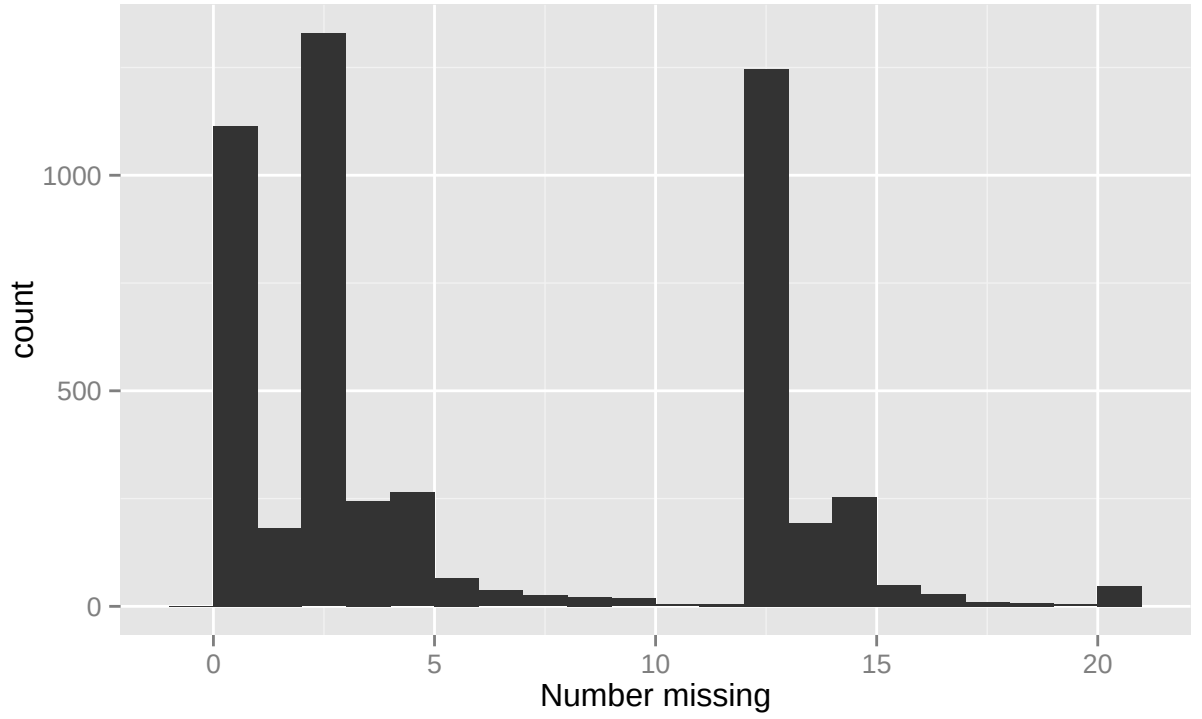

Figure 4: number of missing values for each student (after selecting 20 variables to build a classifier).

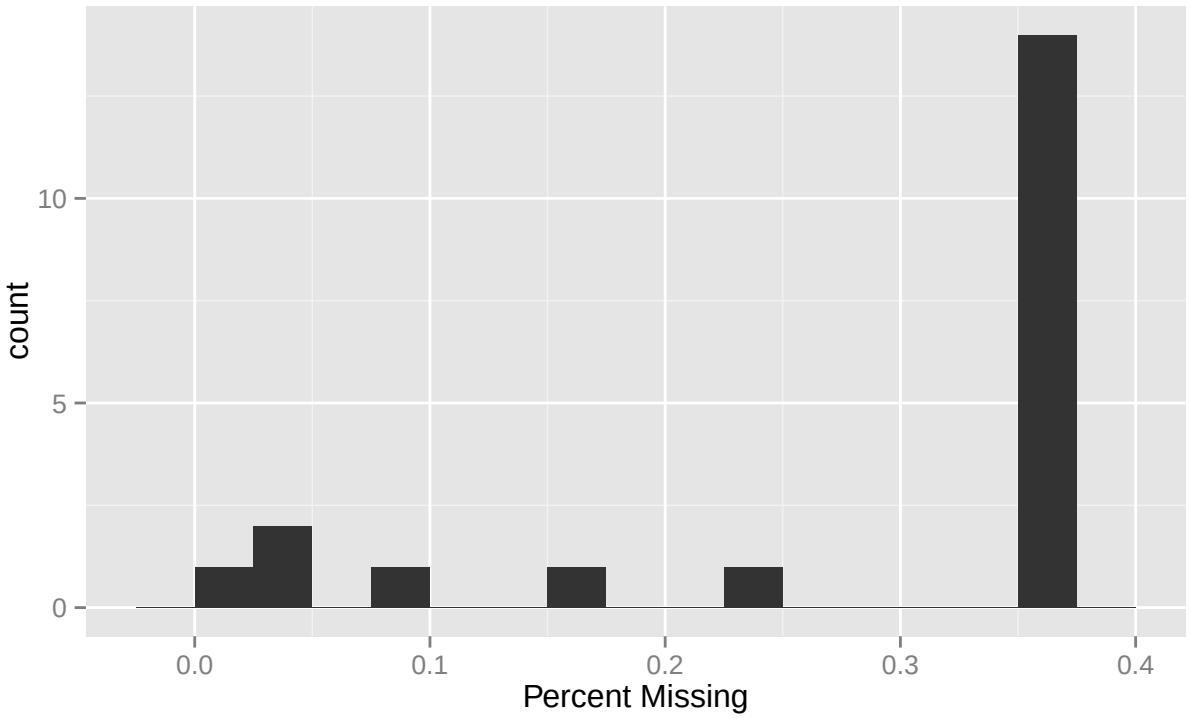

Figure 5: Histogram of the percentage of missing values within each of the 20 selected predictor variables. Most variables are missing about a third of their values.

3. Impute the remaining missing values with nearest neighbor imputation on the 20 predictor variables. I use the `knnImputation` function in the `DMwR` (Torgo 2010) package (setting the number of neighbors to 10).

Exploring the imputed data

Figure 7 shows histograms of all the predictor variables. The nearest-neighbor imputation smoothed out many of the values so that there is some appearance of a continuous scale. Still, the original variables were discrete, so we see 2 to 4 peaks for each histogram. Model-based clustering with a mixture of normal distributions would not be prudent for this data.

Figure 9 plots each predictor variable versus student performance (high or low). All of the variables are associated with student performance, and for some variables, performance-specific interquartile regions do not overlap. Figure 8 shows the pairwise correlations among each of the predictor variables. Most correlations are acceptably close to zero. It seems reasonable to keep all 20 predictor variables.

20 predictor variables can be cumbersome all together, so I used multidimensional scaling to see how the data separate. The `mds` function in the `vegan` package stalled when I tried to use the full dataset, so I took a random subset of 250 students. I plot the MDS components in Figure 10, and I color the points by student performance. The data almost separate by student performance. However, the separation is not complete, and in fact, little to no natural separation or clustering in the data is visible from this plot.

Plan for further work

- Run a basic clustering analysis on the 20 predictor variables in the imputed data. I will use kmeans and hierarchical clustering, and I will determine how much information about student performance

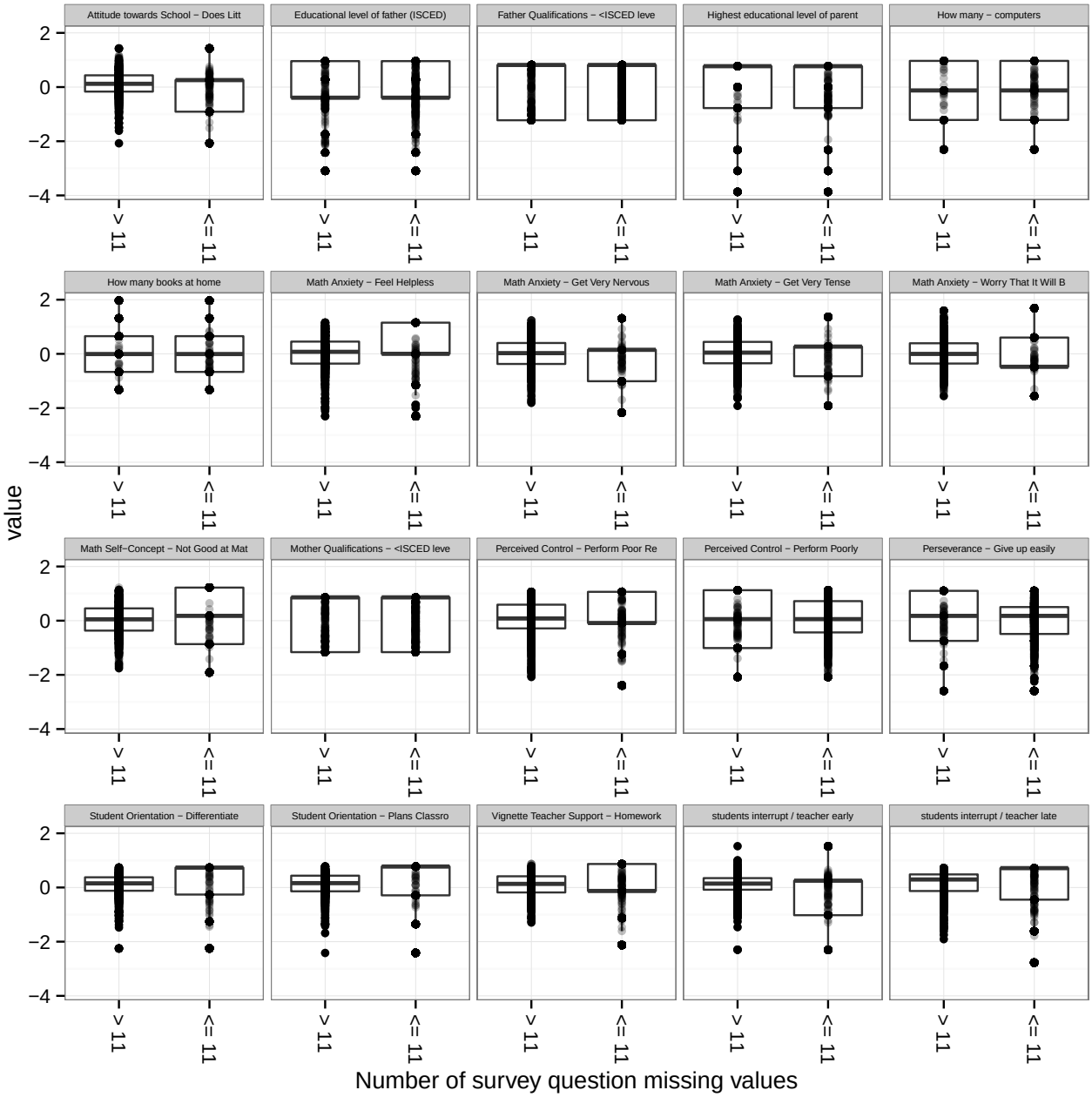

Figure 6: Before imputation, I removed all the students with more than 15 missing values. The purpose of this plot is to determine if I should have reduced this per-student missing value threshold to 11. In each panel, I plot a predictor variable for students with more than 11 missing values and for students with less than 12 missing values. For each variable, the ranges of the variables are nearly the same, and most of the corresponding quartiles are nearly the same. Hence, the imputation succeeded, and I do not need to remove any more students.

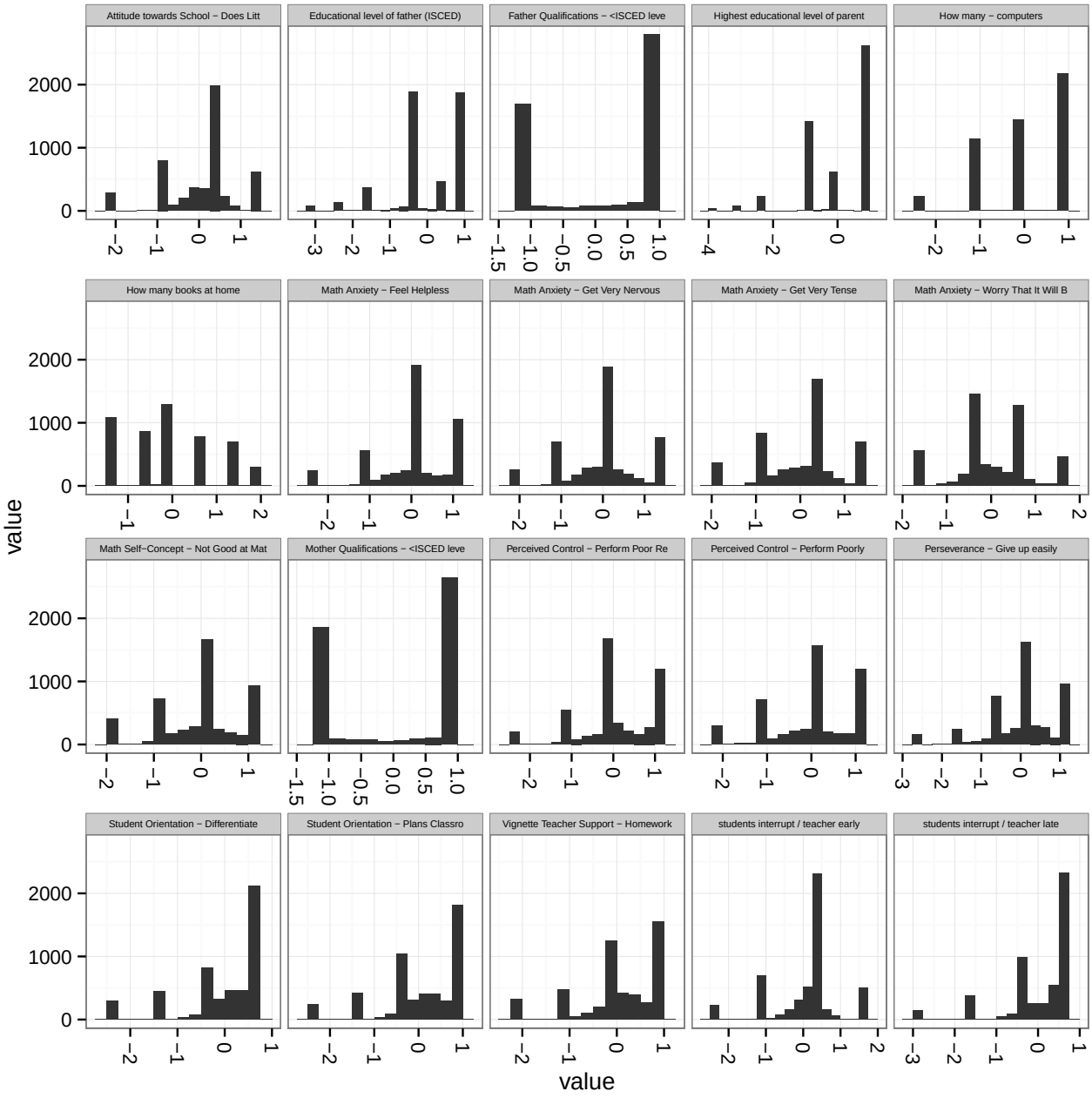

Figure 7: histograms of all the predictor variables. The nearest-neighbor imputation smoothed out many of the values so that there is some appearance of a continuous scale. Still, the original variables were discrete, so we see 2 to 4 peaks for each histogram. Model-based clustering with a mixture of normal distributions would not be prudent for this data.

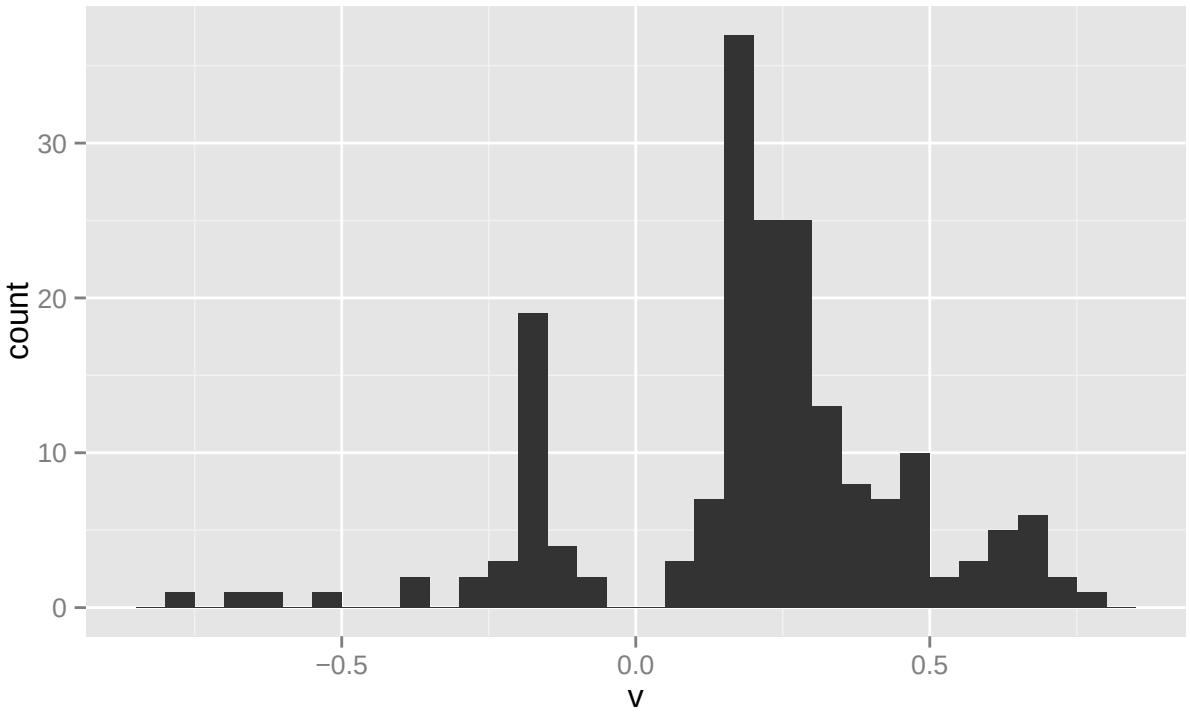

Figure 8: pairwise correlations among each of the predictor variables. Most correlations are acceptably close to zero.

these techniques recover.

- Build a classifier on the preprocessed and imputed data using:
  - Logistic regression
  - Neural networks
  - Random forests
  - Nearest neighbors classification

Acknowledgements

I would like to thank Dr. Cook for steering me in the right direction. The PISA data is messy and cumbersome, and the guidance is very appreciated.

References

"Organization for Economic Co-operation and Development." 2015. http://www.oecd.org/.

Torgo, L. 2010. *Data Mining with R, Learning with Case Studies*. Chapman; Hall/CRC. http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR.

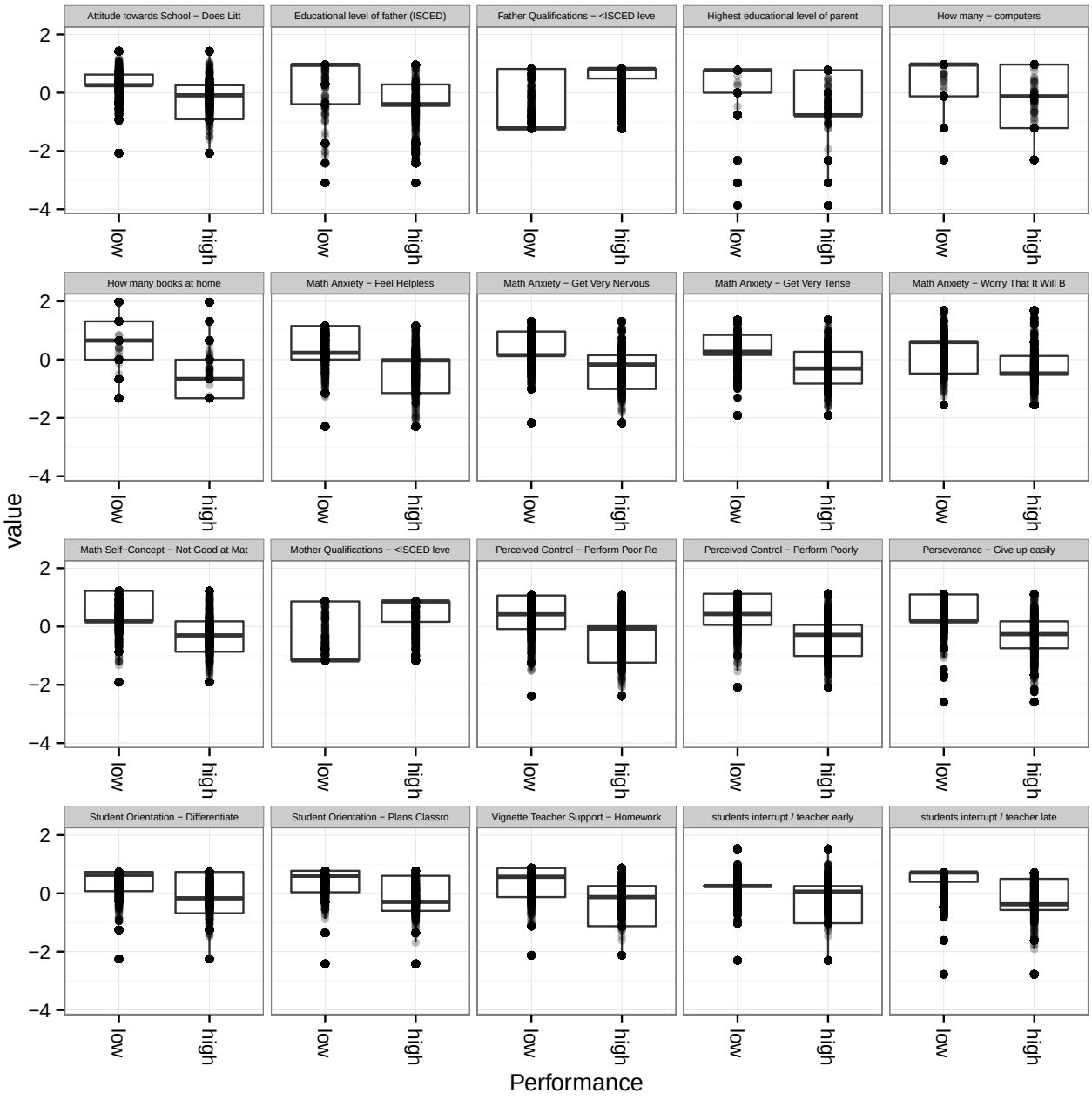

Figure 9: plots of each predictor variable versus student performance (high or low). All of the variables are associated with student performance, and for some variables, performance-specific interquartile regions do not overlap.

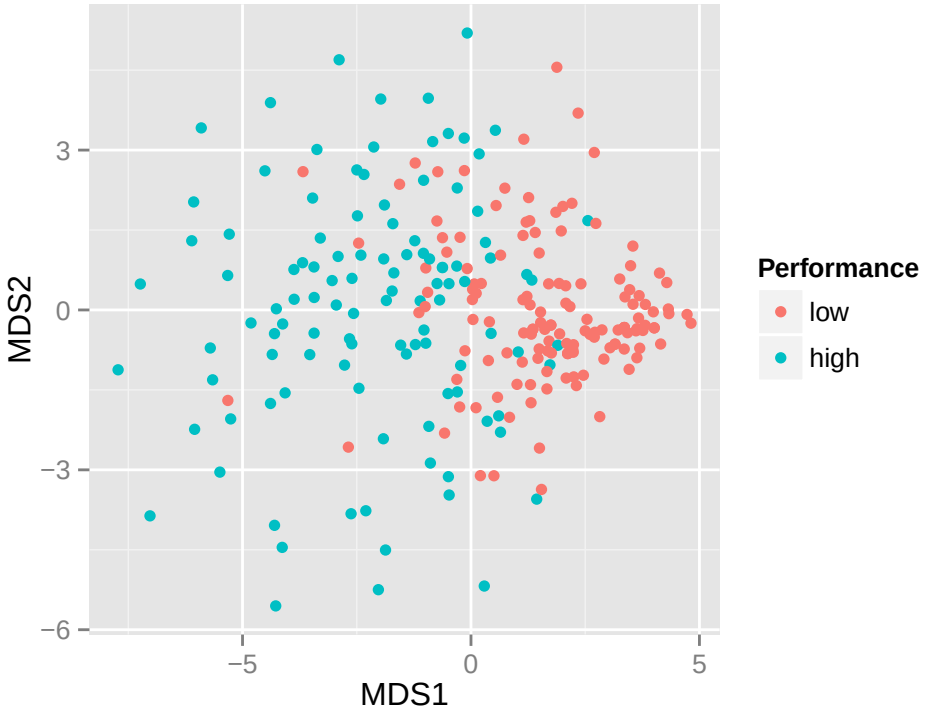

Figure 10: MDS components, where points (students) are colored by performance (high or low). The data almost separate by student performance. However, the separation is not complete, and in fact, little to no natural separation or clustering in the data is visible from this plot.