Sean Deery HW 2

# Introduction

Quality of education affects all people so everyone involved must be aware of how well schools are performing. Education affects students’ future careers, financial stability, physical health, and emotional health. Also, the cost of attending school is significant, so parents and students need to know the quality of schools while deciding on what school to attend.

Beyond the people receiving the education, governments need to know how schools are performing to help with planning for public schools and funding agencies need to know how schools are performing to make lending decisions. Being able to compare schools is essential for society to be able to provide and acquire a high-quality education.

# Analysis and Models

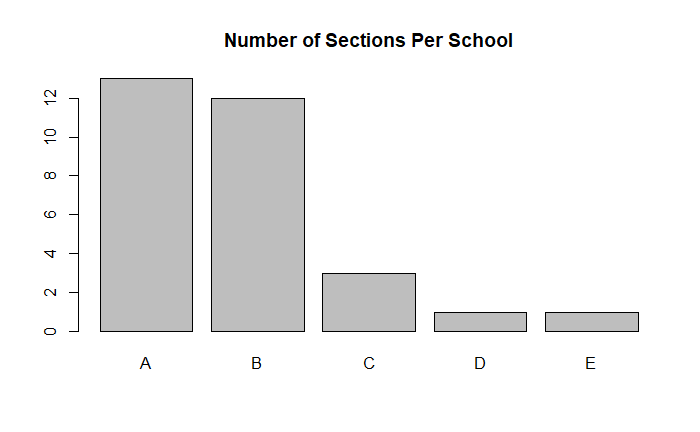
## About the Data

The dataset contains sections of a math course at multiple different schools and includes the number of students who have progressed at different levels. There are eight variables including School, Section, Very Ahead, Middling, Behind, More Behind, Very Behind, and Completed. Out of the level categories, two variables are generally ahead and three are generally behind, so the data may be lopsided toward students being generally behind. There are no missing values in the dataset.

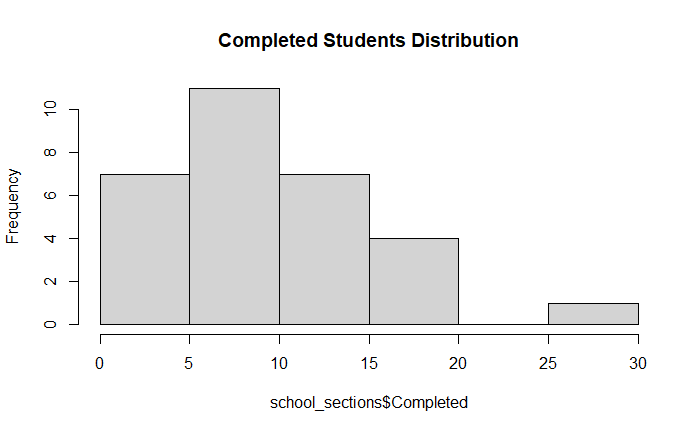
The ‘School’ variable is read in as a character data type which is converted to a factor. The ‘Section’ variable is read in as an integer data type, which is converted to a factor. The rest of the variables, which all represent counts of students, are integers and are not converted to another data type.

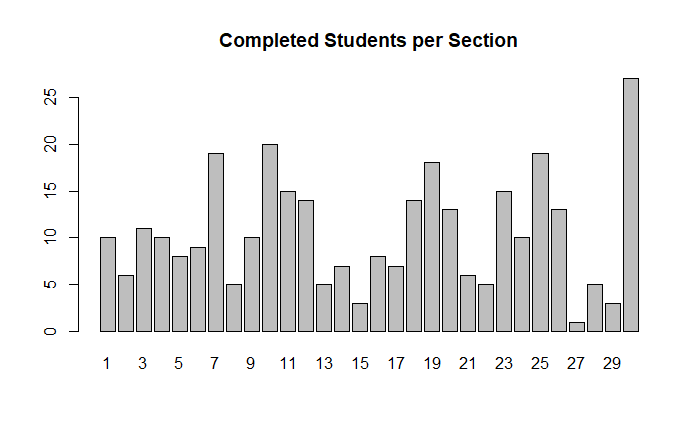
To organize the dataset to be more readable, the variable ‘Section’ is swapped with ‘School’, and the variable ‘Completed’ is moved from the end of the table before ‘Very Ahead’ where it logically belongs.

Looking at the number of ‘Sections’ per ‘School’, schools A and B have over 10 sections, whereas schools C, D, and E have between 1 and 3 sections. This means our results for schools A and B are likely to be more accurate to the truth than the results for schools C, D, and E.



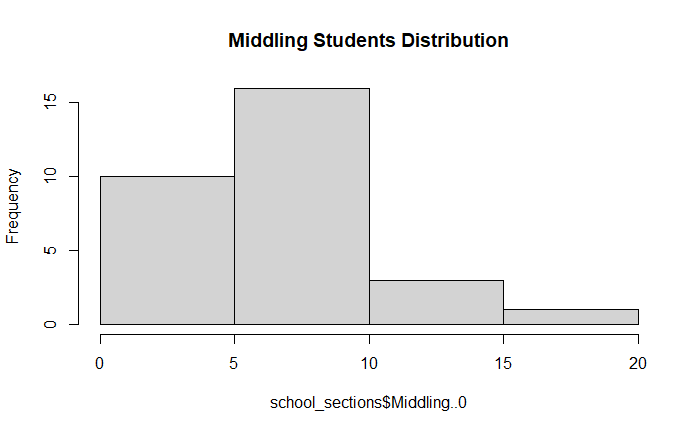
The ‘Completed’ variable lists the number of students for each section who are finished with the course. The histogram shows that ‘Completed’ has a normal distribution that is skewed to the right, with most sections having between 5 and 10 students who have completed the course. The summary statistics show the mean is 10.5 completed students per section, and the median is 10 completed students per section. This further supports that there are outliers with some sections having significantly more students. The bar plot shows that section 1 in school E (row 30), has significantly more completed students than the rest of the sections with 27 students completed. This section also has 56 students who are behind, so this large number may just be due to the size rather than the quality of the section.

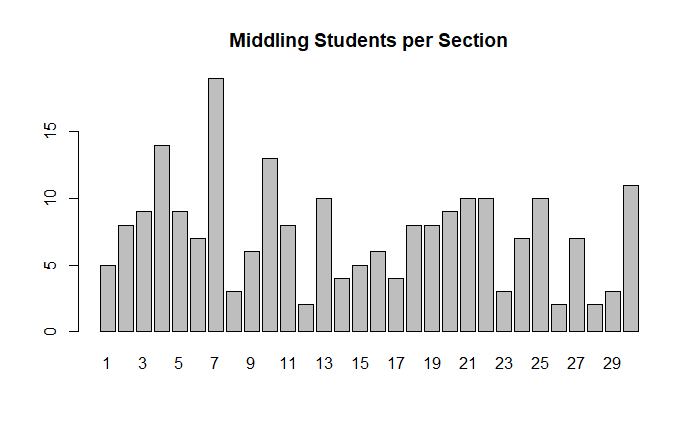




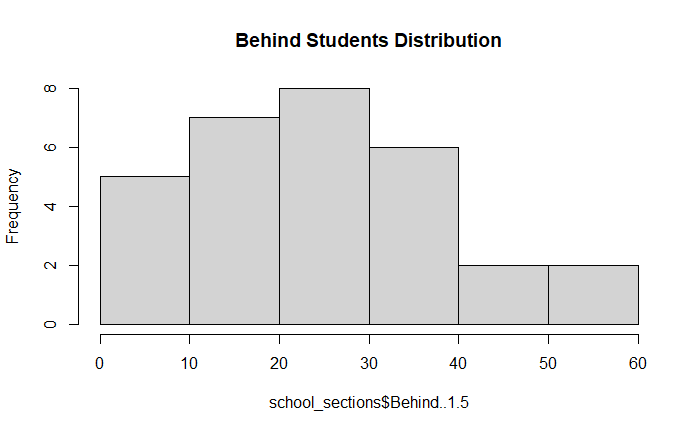
The ‘Very Ahead’ variable lists the number of students for each section who are more than 5 lessons ahead. None of the sections have any students who are in this category.

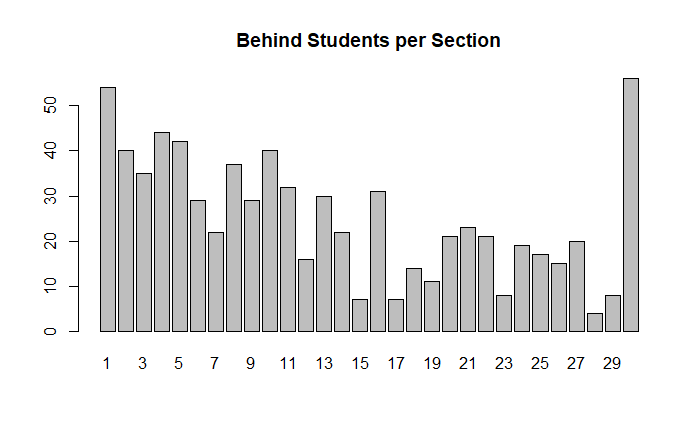
The ‘Middling’ variable lists the number of students for each section who are 0 to 5 lessons ahead. The histogram shows that ‘Middling’ has a normal distribution with a mean of 7.4 and a median of 7.5. The bar plot shows that section 7 in school A (row 7) stands out with 19 students who are middling.



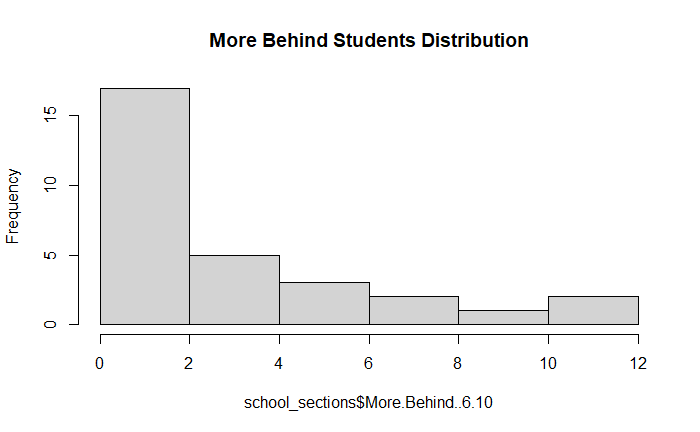


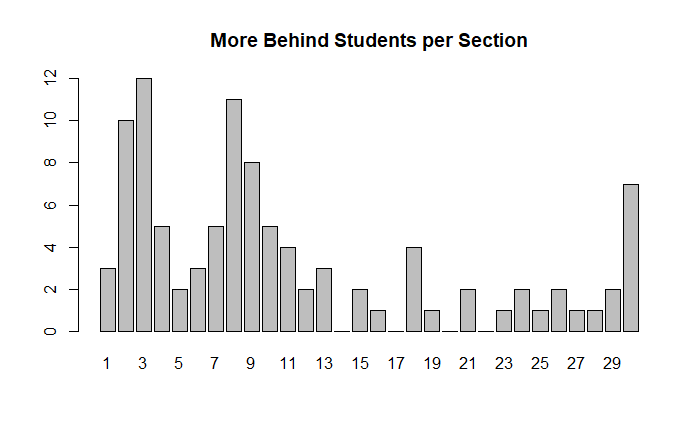
The ‘Behind’ variable lists the number of students for each section who are 1 to 5 lessons behind. The histogram shows ‘Behind’ has a normal distribution skewed to the right with a mean of 25 and a median of 22. The bar plot shows that section 1 in school E (row 30) and section 1 in school A (row 1) have relatively high counts of students behind compared to the rest of the sections.



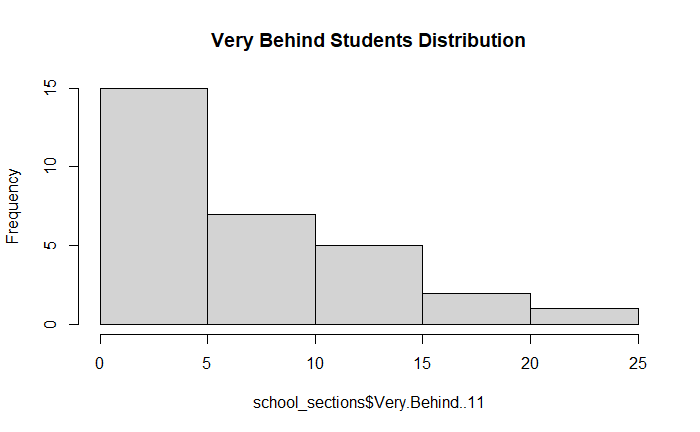


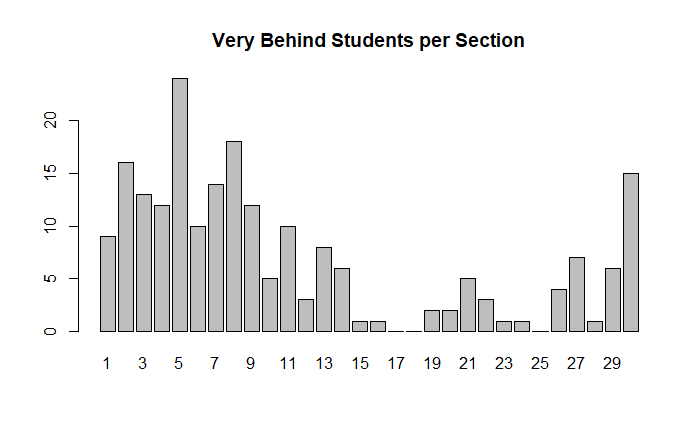
The ‘More Behind’ variable lists the number of students for each section who are 6 to 10 lessons behind. The histogram shows ‘More Behind’ has a Poisson distribution with mostly lower values between 0 and 2. The mean is 3.3 and the median is 2.0. The bar plot shows that there is a small group of sections with 8-12 students who are ‘More Behind’, with most sections having below 6 students who are ‘More Behind’.



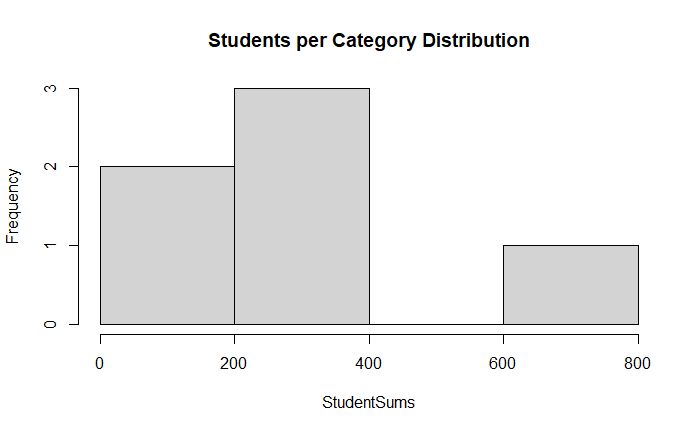


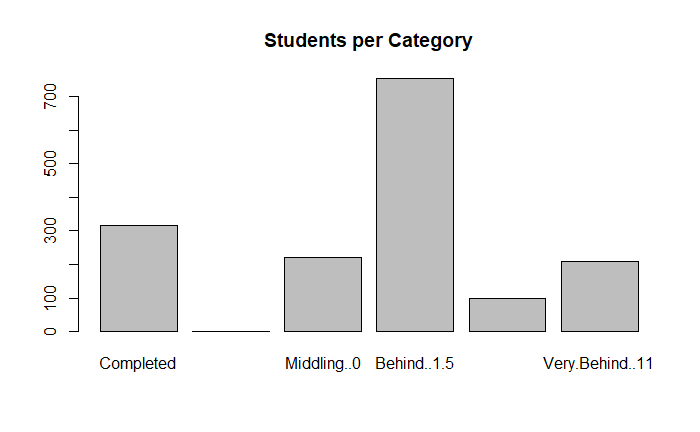
The ‘Very Behind’ variable lists the number of students for each section who are more than 10 lessons behind. The histogram shows ‘Very Behind’ has a Poisson distribution with mostly lower values between 0 and 5. The mean is 7 and the median is 5.5. The bar plot shows section 5 in school A (row 5) has a significantly higher amount of students who are very behind than the rest of the sections.



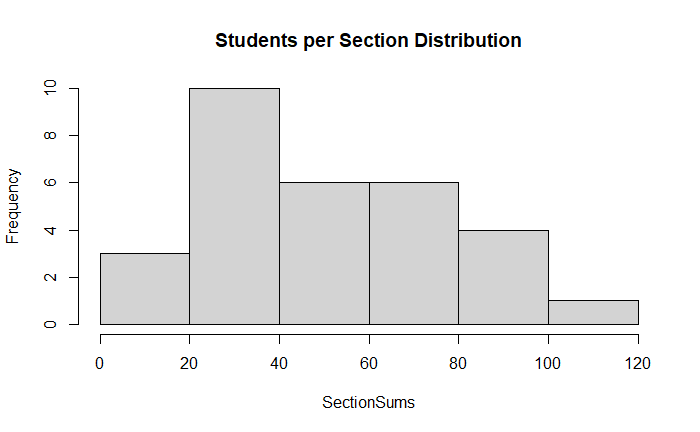


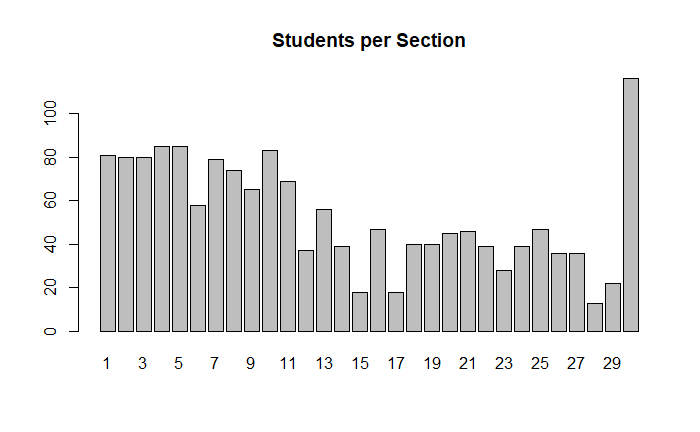
The total number of students per category is calculated by finding the sum of each category column. The histogram shows the totals per category are normally distributed around 300 and are skewed to the right. The mean is 267 and the median is 216, which further supports that there is an outlier on the high end. The bar plot shows that the number of students who are ‘Behind’ is much larger than in any other category.



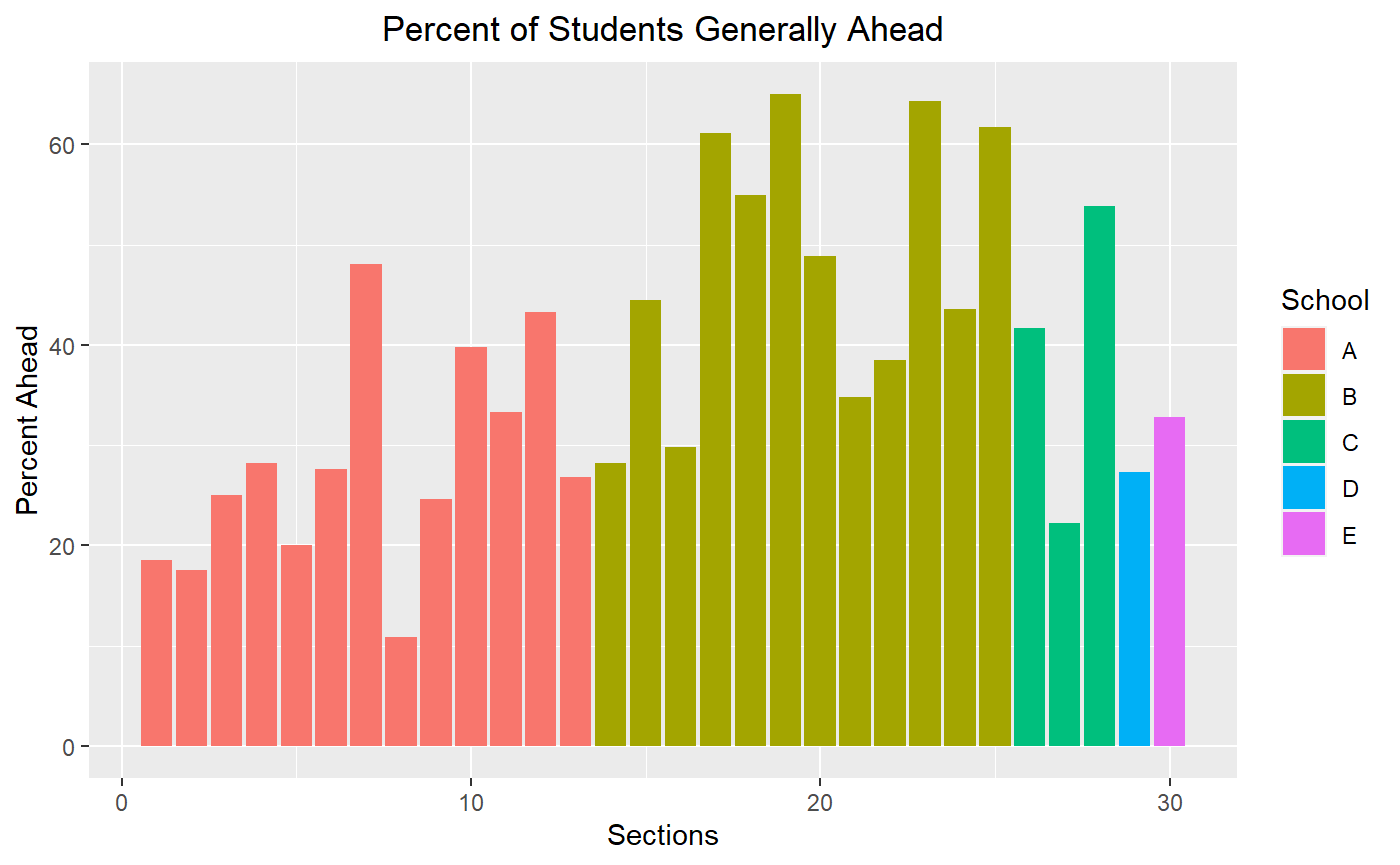


The total number of students per section is calculated by finding the sum of each row. The histogram shows the totals per section are normally distributed around 40 and are skewed to the right. The mean is 53 and the median is 46, which further supports that there is an outlier on the high end. The bar plot shows that section 1 in school E (row 30) has a much higher number of students than the rest of the sections.





To compare the number of students per section who are generally ahead versus generally behind, the counts for the categories are turned into percentages to get them on the same scale. Then the students who are ‘Completed’, ‘Very Ahead’, and ‘Middling’ are combined into ‘Generally\_Ahead\_Percent’ and the students who are ‘Behind’, ‘More Behind’, and ‘Very Behind’ are combined into ‘Generally\_Behind\_Percent’. The barplot shows that the sections in school B outperformed the rest of the schools. The table shows the percentages aggregated by school, which further supports school B as the best-performing school with 47 percent of students who are generally ahead.



A picture containing table

Description automatically generated

# Conclusions

The results show that school B is the best-performing school. It has the highest percentage of students who are on track or ahead of schedule for the math course. It also has a relatively high number of sections, which strengthens the conclusion since the school was able to perform well in multiple sections rather than just one.

School C is the second-best performing school. It has the second-highest percentage of students who are on track or ahead of schedule for the math course. The school only has three sections, but two of them had over 40% of their students on track or ahead of schedule.

School E shows up as the third-best performing school when looking at the percentage of students who are on track or ahead of schedule. It only has one section to analyze, but that section has significantly more students than the rest of the sections.