**Kendra Danforth Osburn**

**IST 707 HW2**

**January 23, 2019**

*R Version*

**Introduction**

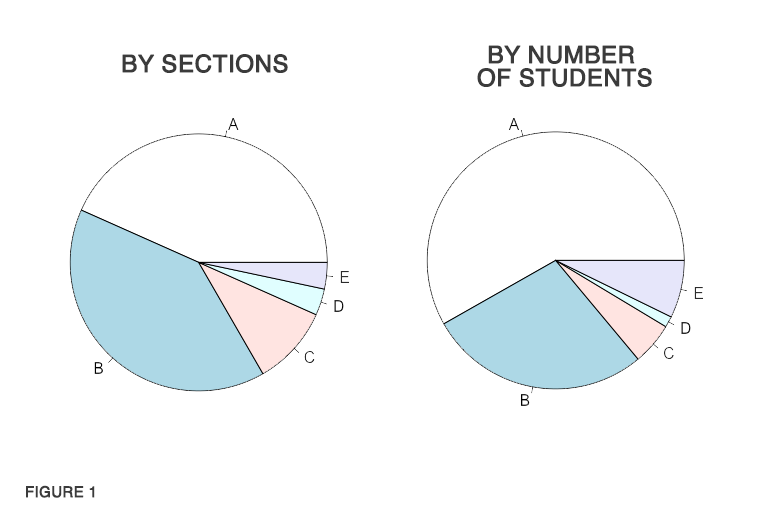
Education is among the most significant challenges facing our society. Yet despite near-universal agreement on the importance of a quality education, the debate over how to achieve that goal rages on. At the heart of this debate is the problem of defining and measuring educational outcomes. Culture, quality, resources, and methods differ widely from city to city, school to school, even teacher to teacher. As a result, students taking the same classes with similar material may have a wide range of experiences and outcomes. Nevertheless, many of the educational policy decisions that impact schools are made at the state and federal level. With that in mind, it’s crucial that educators and researchers create an objective set of standards to evaluate students and schools.

In approaching the issue of education from a data science perspective, three major questions must be answered. First: how are students performing? Which types of assessments are most meaningful in gauging achievement, and what topics or skills do they cover? Second: which schools have the highest- and lowest-performing students, and what are their characteristics? Finally: what conclusions can we draw about the methods and characteristics of high performing schools? From funding, to class size, to technology, experts have proposed a variety of factors that have an impact on education quality. By analyzing both students and schools, researchers can determine which of these factors contribute most strongly to a quality education.

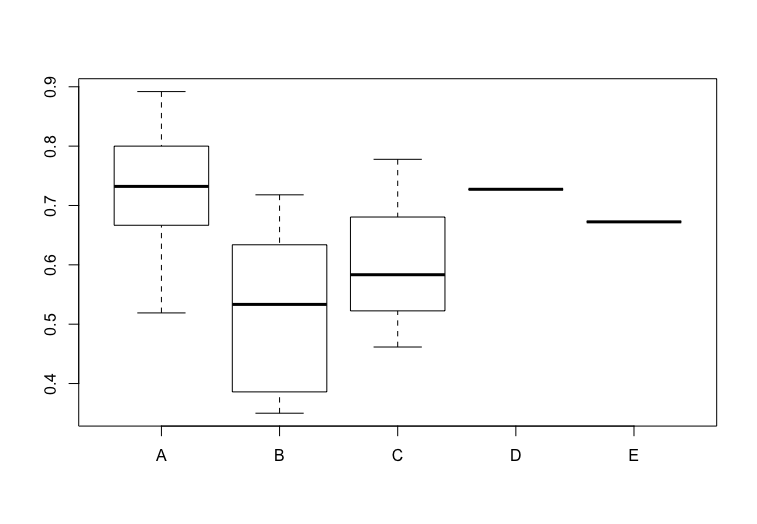
**Analysis and Models**

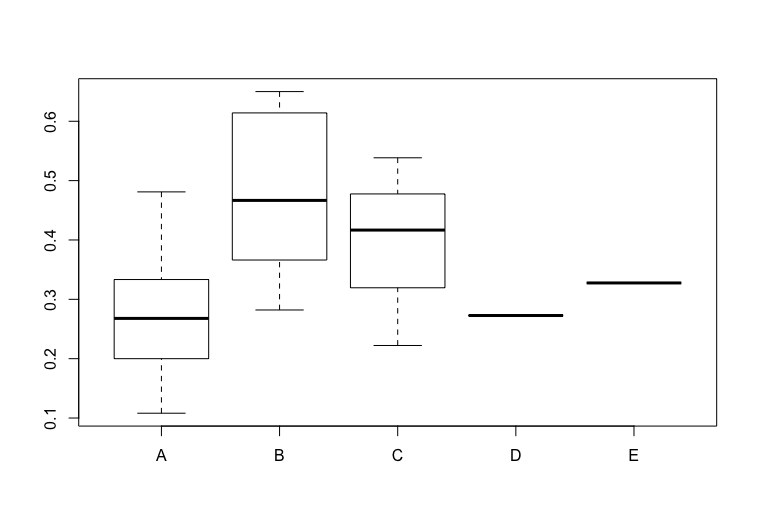
This dataset covers 1601 students in 30 sections of a particular math class across 5 schools. School A has 13 sections and 932 students. School B has 12 sections and 446 students. School C has 3 sections and 85 students. Schools D and E each have one section and 22 and 116 students, respectively. There is no indication where these schools are located, what grades these students are in, if one section is equal to one classroom or if a section can have many classrooms or teachers. The only information given is school (A, B, C, D or E), section (1-13), and the number of students per section who have completed the course, who are very ahead of the course (defined as 5 lessons ahead), who are in the middle of the course (0-5 lessons ahead), who are behind in the course (1-5 lessons behind), who are more behind in the course (6-10 lessons behind) and those who are very behind in the course (more than 10 lessons behind).

In FIGURE 1, the “BY SECTIONS” pie chart is how the data looked in the beginning -- segmented by section. From this pie chart, it’s reasonable to see that school A has 13 sections, B has 12 sections, C has 3 sections and D and E each have one section. But what is a section? This information is not known. Is there a standard number of students per section? No, there is not. In FIGURE 1 the “BY NUMBER OF STUDENTS” pie chart shows the distribution of students across each school. Schools C, D & E are small enough that they could potentially be combined into O for “Other.”

****

The first boxplot was created by summing all of the students who were “behind” (“behind,” “more behind,” “very behind”) and plotting those percentages per school. From this plot, it’s clear that the majority of School A is behind in this course. It is equally evident that School B is doing considerably better keeping up with the coursework. The second boxplot was created by summing all the students who were “not behind” (“completed,” “very ahead”, or “middling,”). As this boxplot is literally representing the inverse of the first boxplot, it clearly affirms what we saw in the first boxplot, School B is performing better than School A (assuming the metric of performance is being “not behind” in this math course).

****

****

**Results**

The data was analyzed over three days and 21 hours. It was divided by school, by section, by number of students, by percentage of students in relation to the students studied as a whole, as well as many other less informative metrics. The data was diagramed, box-plotted, pie-charted, converted, unconverted and reconverted. In the end, the results are clear -- more data is needed. With this limited data and tight time constraint, the only clear story is that School B, the school with equal sections but half as many students as School A, is overall more current (not “behind”) with this coursework.

**Conclusions**

Among the schools included in this dataset, students from smaller schools and sections generally performed better when compared to students from larger schools and larger sections. Based on this result, it is tempting to suggest that smaller class sizes offer a more positive learning environment. Nevertheless, many questions about this dataset remain unanswered.

First, the dataset’s measures student achievement based on “lessons completed.” How does this course define a lesson? Is it a written assignment completed at home, a test completed in the classroom, or some combination of tasks? Next, there are questions about the sections and schools themselves. How does the dataset define a “section”? Is it a class, or some other unit? This brings us to the issue of teachers. Which teachers are responsible for which sections? Can they teach multiple sections, or just one? With answers to these questions, one may find that student achievement is actually more closely tied to teacher than it is to class size.

Finally, there are questions about the data collection process. How was the data gathered? Was this process reliable? For instance, why do schools D & E have only one section each of this math course? Why does school E have so many students included in just one section? Was the data recorded incorrectly? To some extent, combining these schools into an “Other” category may help mitigate issues of data quality and provide more meaningful results. But given the questions discussed above, and without more insight into the data collection methodology, it’s difficult to present any findings with confidence.

*Note: As I wrote in my lengthy email, I got very lost and tangled in this assignment. I am excited to grow both in the process of coding/writing about these exercises as well as my time management. I am not at all pleased with my middle sections as they kept changing as I changed my approach. I look forward to learning how I can mitigate these (many) issues in the future.*